Minimax-optimal rates for sparse additive models over kernel classes via convex programming

Garvesh Raskutti[1] Martin J. Wainwright[1,2]
garveshr@stat.berkeley.edu wainwrig@stat.berkeley.edu

Bin Yu[1,2]
binyu@stat.berkeley.edu

Departments of Statistics[1], and EECS[2]
UC Berkeley, Berkeley, CA 94720

Abstract

Sparse additive models are families of d-variate functions that have the additive decomposition $f^* = \sum_{j \in S} f_j^*$, where S is a unknown subset of cardinality $s \ll d$. We consider the case where each component function f_j^* lies in a reproducing kernel Hilbert space, and analyze a simple kernel-based convex program for estimating the unknown function f^*. Working within a high-dimensional framework that allows both the dimension d and sparsity s to scale, we derive convergence rates in the $L^2(\mathbb{P})$ and $L^2(\mathbb{P}_n)$ norms. These rates consist of two terms: a *subset selection term* of the order $\frac{s \log d}{n}$, corresponding to the difficulty of finding the unknown s-sized subset, and an *estimation error* term of the order $s\,\nu_n^2$, where ν_n^2 is the optimal rate for estimating an univariate function within the RKHS. We complement these achievable results by deriving minimax lower bounds on the $L^2(\mathbb{P})$ error, thereby showing that our method is optimal up to constant factors for sub-linear sparsity $s = o(d)$. Thus, we obtain optimal minimax rates for many interesting classes of sparse additive models, including polynomials, splines, finite-rank kernel classes, as well as Sobolev smoothness classes.

1 Introduction

The past decade has witnessed a flurry of research on sparsity constraints in statistical models. Sparsity is an attractive assumption for both practical and theoretical reasons: it leads to more interpretable models, reduces computational cost, and allows for model identifiability even under high-dimensional scaling, where the dimension d exceeds the sample size n. While a large body of work has focused on sparse linear models, many applications call for the additional flexibility provided by non-parametric models. In the general setting, a non-parametric regression model takes the form $y = f^*(x_1, \ldots, x_d) + w$, where $f : \mathbb{R}^d \to \mathbb{R}$ is the unknown regression function, and w is scalar observation noise. Unfortunately, this general non-parametric model is known to suffer severely from the so-called "curse of dimensionality", in that for most natural function classes (e.g., twice differentiable functions), the sample size n required to achieve any given error grows exponentially in the dimension d.

Given this curse of dimensionality, it is essential to further limit the complexity of possible functions f^*. One attractive candidate are the class of *additive non-parametric models* [15], in which the function f^* has an additive decomposition of the form

$$f^*(x_1, x_2, \ldots, x_d) = \sum_{j=1}^{d} f_j^*(x_j), \tag{1}$$

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
AUG 2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE
Minimax-optimal rates for sparse additive models over kernel classes via convex programming

5a. CONTRACT NUMBER
5b. GRANT NUMBER
5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER
5e. TASK NUMBER
5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University California, Berkeley,Departments of Statistics,Berkeley,CA,94720

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)
11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Sparse additive models are families of d-variate functions that have the additive decomposition f = P j2S f j , where S is a unknown subset of cardinality s d. We consider the case where each component function f j lies in a reproducing kernel Hilbert space, and analyze a simple kernel-based convex program for estimating the unknown function f . Working within a high-dimensional framework that allows both the dimension d and sparsity s to scale, we derive convergence rates in the L2(P) and L2(Pn) norms. These rates consist of two terms: a subset selection term of the order s log d n , corresponding to the di culty of nding the unknown s-sized subset, and an estimation error term of the order s 2 n, where 2 n is the optimal rate for estimating an univariate function within the RKHS. We complement these achievable results by deriving minimax lower bounds on the L2(P) error, thereby showing that our method is optimal up to constant factors for sub-linear sparsity s = o(d). Thus, we obtain optimal minimax rates for many interesting classes of sparse additive models, including polynomials, splines, nite-rank kernel classes, as well as Sobolev smoothness classes.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**35**	

where each component function f_j^* is univariate. Given this decoupling, this function class no longer suffers from the exponential explosion in sample size of the general non-parametric model. Nonetheless, one still requires a sample size $n \gg d$ for consistent estimation; note that this is true even for the linear model, which is a special case of equation (1).

A natural extension is the class of *sparse additive models*, in which the unknown regression function is assumed to have a decomposition of the form

$$f^*(x_1, x_2 \ldots, x_d) = \sum_{j \in S} f_j^*(x_j), \tag{2}$$

where $S \subseteq \{1, 2, \ldots, d\}$ is some unknown subset of cardinality $|S| = s$. Of primary interest is the case when the decomposition is genuinely sparse, so that $s \ll d$. To the best of our knowledge, this model class was first introduced in Lin and Zhang [21], and has since been studied by various researchers (e.g., [17, 23, 28, 37]). Note that the sparse additive model (2) is a natural generalization of the sparse linear model, to which it reduces when each univariate function is constrained to be linear.

In past work, several groups have proposed computationally efficient methods for estimating sparse additive models (2). Just as ℓ_1-based relaxations such as the Lasso have desirable properties for sparse parametric models, more general ℓ_1-based approaches have proven to be successful in this setting. Lin and Zhang [21] proposed the COSSO method, which extends the Lasso to cases where the component functions f_j^* lie in a reproducing kernel Hilbert space (RKHS); see also Yuan [37] for a similar extension of the non-negative garrote [7]. Bach [3] analyzes a closely related method for the RKHS setting, in which least-squares loss is penalized by an ℓ_1-sum of Hilbert norms, and establishes consistency results in the classical (fixed d) setting. Other related ℓ_1-based methods have been proposed in independent work by Koltchinskii and Yuan [17], Ravikumar et al. [28] and Meier et al. [23], and analyzed under high-dimensional scaling. As we describe in more detail in Section 3.3, each of the above papers establish consistency and convergence rates for the prediction error under certain conditions on the covariates as well as the sparsity s and dimension d. However, it is not clear whether the rates obtained in these papers are sharp for the given methods, nor whether the rates are minimax-optimal.

This paper makes two main contributions to this on-going line of research. Our first contribution is to analyze a simple polynomial-time method for estimating sparse additive models and provide upper bounds on the error in both the $L^2(\mathbb{P}_n)$ and $L^2(\mathbb{P})$ norms. Our method is based on a combination of least-squares loss with two ℓ_1-based sparsity penalty terms, one corresponding to an $\ell_1/L^2(\mathbb{P}_n)$ norm and the other an $\ell_1/\|\cdot\|_{\mathcal{H}}$ norm. This combination yields a second-order cone program, for which solutions can be computed in polynomial time using interior-point methods (see §4, 11 in Boyd and Vandenberghe [6] for details). Although closely related to the methods considered in past work [3, 17, 23, 28], our estimator differs in the particular form of regularization, and we suspect that these differences are important in obtaining optimal convergence rates. Our first main result (Theorem 1) shows that that with high probability, the error of our procedure, in either the squared $L^2(\mathbb{P}_n)$ or $L^2(\mathbb{P})$ norms, is bounded by $\mathcal{O}\big(\frac{s \log d}{n} + s\nu_n^2\big)$. Each of these two terms has a natural interpretation. The quantity $\frac{s \log d}{n}$ is a *subset selection term*, which reflects the difficulty of extracting the s-sized subset of active functions from the total d. On the other hand, the quantity ν_n^2 corresponds to the optimal rate for estimating a single univariate function, so that $s\nu_n^2$ corresponds to the *s-dimensional estimation error* associated with estimating s univariate functions. This latter term depends on the sparsity s but *not* on the ambient dimension d. In order

to illustrate these rates more concretely, we discuss two particular consequences of Theorem 1. First, Corollary 1 applies to parametric function classes and m-rank kernel classes, where $\nu_n^2 \sim \frac{m}{n}$. Second, Corollary 2 applies to various types of non-parametric classes, among them Sobolev spaces, where $\nu_n^2 \sim n^{-2\alpha/(2\alpha+1)}$, for some $\alpha > 1/2$.

Our second contribution is complementary in nature, in that it establishes lower bounds that hold uniformly over all algorithms. These minimax lower bounds, stated in Theorem 2, are specified in terms of the metric entropy of the underlying univariate function classes. For both finite-rank kernel classes and Sobolev-type classes, these lower bounds match our achievable results, as stated in Corollaries 1 and 2, up to constant factors in the regime of sub-linear sparsity ($s = o(d)$). Thus, for these function classes, we have a sharp characterization of the associated minimax rates. The proofs of these results are based on characterizing the packing entropies of the class of sparse additive models, combined with the use of the Fano method.

The lower bounds derived in this paper initially appeared in the Proceedings of the NIPS Conference (December 2009). As we were completing this write-up, we became aware of concurrent work by Koltchinskii and Yuan [18] (hereafter KY) that analyzes essentially the same estimator as that used to prove upper bounds in this paper. As with our analysis, they assume that the unit ball of each univariate Hilbert class $\mathcal{H}_j$, for $j = 1, \ldots, d$, is bounded. Under this assumption, they derive a result (Theorem 3 in their paper) that contains the two terms involved in our Theorem 1, but also includes additional pre-factors that depend on a *global bound* on the function class—that is, the quantity $C(\mathcal{F}_{d,s}) = \sup_{f \in \mathcal{F}_{d,s}} \|f\|_\infty$, where $\mathcal{F}_{d,s}$ is the class of s-sparse additive models in d dimensions. Our result (Theorem 1 in our paper) requires only that each univariate function is bounded, which is much less restrictive than global boundedness. If the quantity $C(\mathcal{F}_{d,s})$ remains bounded independently of the dimension d and sparsity s, then their result matches our rate up to constant factors. On the other hand, if $C(\mathcal{F}_{d,s})$ scales with (d, s), then our bound, which has no dependence on this quantity, is tighter. It is worth noting that the condition $C(\mathcal{F}_{d,s}) = O(1)$—an assumption that might seem innocuous at first sight—can be fairly restrictive for sparse additive models under the high-dimensional scaling $(d, s) \to +\infty$. If a global boundedness condition is imposed, the rates are not minimax-optimal in general—for instance, see Example 1 in Section 3.3. In addition to global boundedness, there are other differences between the two papers. For instance, they analyze a slightly more general class of quadratic-type losses, as opposed to the least-squares loss considered here, and their analysis involves directly imposing RIP conditions on fixed design matrices, whereas we consider the case of random design with independent co-ordinates (although our results hold albeit with slightly worse constants if we impose RIP conditions instead of independence).

The remainder of the paper is organized as follows. In Section 2, we provide background on kernel spaces and the class of sparse additive models considered in this paper. Section 3 is devoted to the statement of our main results and discussion of their consequences; it includes description of our method, the convergence rates that it achieves, and a matching set of minimax lower bounds. Section 4 is devoted the proofs of our upper and lower bounds, presented in Sections 4.1 and Section 4.2 respectively, with the more technical details deferred to the Appendices. We conclude with a discussion in Section 5.

2 Background and problem set-up

We begin with some background on reproducing kernel Hilbert spaces, before providing a precise definition of the class of sparse additive models studied in this paper.

2.1 Reproducing kernel Hilbert spaces

Given a subset $\mathcal{X} \subset \mathbb{R}$ and a probability measure $\mathbb{Q}$ on $\mathcal{X}$, we consider a Hilbert space $\mathcal{H} \subset L^2(\mathbb{Q})$, meaning a family of functions $g : \mathcal{X} \to \mathbb{R}$, with $\|g\|_{L^2(\mathbb{Q})} < \infty$, and an associated inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ under which $\mathcal{H}$ is complete. The space $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) if there exists a symmetric function $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ such that: (a) for each $x \in \mathcal{X}$, the function $\mathbb{K}(\cdot, x)$ belongs to the Hilbert space $\mathcal{H}$, and (b) we have the reproducing relation $f(x) = \langle f, \mathbb{K}(\cdot, x) \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. Any such kernel function must be positive semidefinite; under suitable regularity conditions, Mercer's theorem [25] guarantees that the kernel has an eigen-expansion of the form

$$\mathbb{K}(x, x') = \sum_{\ell=1}^{\infty} \mu_\ell \phi_\ell(x) \phi_\ell(x'), \tag{3}$$

where $\mu_1 \geq \mu_2 \geq \mu_3 \geq \ldots \geq 0$ are a non-negative sequence of eigenvalues, and $\{\phi_j\}_{j=1}^{\infty}$ are the associated eigenfunctions, taken to be orthonormal in $L^2(\mathbb{Q})$. The decay rate of these eigenvalues will play a crucial role in our analysis, since they ultimately determine the rate ν_n for the univariate RKHS's in our function classes.

Since the eigenfunctions $\{\phi_\ell\}_{\ell=1}^{\infty}$ form an orthonormal basis, any function $f \in \mathcal{H}$ has an expansion of the $f(x) = \sum_{\ell=1}^{\infty} a_\ell \phi_\ell(x)$, where $a_\ell = \langle f, \phi_\ell \rangle_{L^2(\mathbb{Q})} = \int_{\mathcal{X}} f(x) \phi_\ell(x) \, d\mathbb{Q}(x)$ are (generalized) Fourier coefficients. Associated with any two functions in $\mathcal{H}$—say $f = \sum_{\ell=1}^{\infty} a_\ell \phi_\ell$ and $g = \sum_{\ell=1}^{\infty} b_\ell \phi_\ell$—are two distinct inner products. The first is the usual inner product in the space $L^2(\mathbb{Q})$—namely, $\langle f, g \rangle_{L^2(\mathbb{Q})} := \int_{\mathcal{X}} f(x) g(x) \, d\mathbb{Q}(x)$. By Parseval's theorem, it has an equivalent representation in terms of the expansion coefficients—namely

$$\langle f, g \rangle_{L^2(\mathbb{P})} = \sum_{\ell=1}^{\infty} a_\ell b_\ell.$$

The second inner product, denoted $\langle f, g \rangle_{\mathcal{H}}$, is the one that defines the Hilbert space; it can be written in terms of the kernel eigenvalues and generalized Fourier coefficients as

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{a_\ell b_\ell}{\mu_\ell}.$$

For more background on reproducing kernel Hilbert spaces, we refer the reader to various standard references [2, 29, 30, 34, 12].

2.2 Sparse additive models over RKHS

For each $j = 1, \ldots, d$, let $\mathcal{H}_j \subset L^2(\mathbb{Q})$ be a reproducing kernel Hilbert space of univariate functions on the domain $\mathcal{X}$. Without loss of generality (by re-centering the functions as needed), we may assume that

$$\mathbb{E}[f_j(x)] = \int_{\mathcal{X}} f_j(x) d\mathbb{Q}(x) \; = \; 0 \qquad \text{for all } f_j \in \mathcal{H}_j,$$

and for each $j = 1, 2, \ldots, d$. For a given subset $S \subset \{1, 2, \ldots, d\}$, we define

$$\mathcal{H}(S) := \big\{ f = \sum_{j \in S} f_j \;\mid\; f_j \in \mathcal{H}_j, \text{ and } \|f_j\|_{\mathcal{H}_j} \leq 1 \;\forall\, j \in S \big\}, \tag{4}$$

corresponding to the class of functions $f : \mathcal{X}^d \to \mathbb{R}$ that decompose as sums of univariate functions on co-ordinates lying within the set S. Note that $\mathcal{H}(S)$ is also (a subset of) a reproducing kernel Hilbert space, in particular with the norm

$$\|f\|^2_{\mathcal{H}(S)} = \sum_{j \in S} \|f_j\|^2_{\mathcal{H}_j},$$

where $\|\cdot\|_{\mathcal{H}_j}$ denotes the norm on the univariate Hilbert space $\mathcal{H}_j$. Finally, for a cardinality $s \in \{1, 2, \ldots, \lfloor d/2 \rfloor\}$, we define the function class

$$\mathcal{F}_{d,s,\mathcal{H}} := \bigcup_{\substack{S \subset \{1,2,\ldots,d\} \\ |S| = s}} \mathcal{H}(S). \tag{5}$$

To ease notation, we frequently adopt the shorthand $\mathcal{F} = \mathcal{F}_{d,s,\mathcal{H}}$, but the reader should recall that $\mathcal{F}$ depends on the choice of Hilbert spaces $\{\mathcal{H}_j\}_{j=1}^d$, and moreover, that we are actually studying a *sequence of function classes* indexed by (d, s).

Now let $\mathbb{P} = \mathbb{Q}^d$ denote the product measure on the space $\mathcal{X}^d \subseteq \mathbb{R}^d$. Given an arbitrary $f^* \in \mathcal{F}$, we consider the observation model

$$y_i = f^*(x_i) + w_i, \quad \text{for } i = 1, 2, \ldots, n, \tag{6}$$

where $\{w_i\}_{i=1}^n$ is an i.i.d. sequence of standard normal variates, and $\{x_i\}_{i=1}^n$ is a sequence of design points in $\mathbb{R}^d$, sampled in an i.i.d. manner from $\mathbb{P}$.

Given an estimate $\widehat{f}$, our goal is to bound the error $\widehat{f} - f^*$ under two norms. The first is the *usual* $L^2(\mathbb{P})$ *norm* on the space $\mathcal{F}$; given the product structure of $\mathbb{P}$ and the additive nature of any $f \in \mathcal{F}$, it has the additive decomposition $\|f\|^2_{L^2(\mathbb{P})} = \sum_{j=1}^d \|f_j\|^2_{L^2(\mathbb{Q})}$. In addition, we consider the error in the *empirical* $L^2(\mathbb{P}_n)$*-norm* defined by the sample $\{x_i\}_{i=1}^n$, defined as $\|f\|^2_{L^2(\mathbb{P}_n)} := \frac{1}{n}\sum_{i=1}^n f^2(x_i)$. Unlike the $L^2(\mathbb{P})$ norm, this norm does not decouple across the dimensions, but part of our analysis will establish an approximate form of such decoupling. For shorthand, we frequently use the notation $\|f\|_2 = \|f\|_{L^2(\mathbb{P})}$ and $\|f\|_n = \|f\|_{L^2(\mathbb{P}_n)}$ for a d-variate function $f \in \mathcal{F}$. With a minor abuse of notation, for a univariate function $f_j \in \mathcal{H}_j$, we also use the shorthands $\|f_j\|_2 = \|f_j\|_{L^2(\mathbb{Q})}$ and $\|f_j\|_n = \|f\|_{L^2(\mathbb{Q}_n)}$.

3 Main results and their consequences

This section is devoted to the statement of our main results, and discussion of some of their consequences. We begin in Section 3.1 by describing a regularized M-estimator for sparse additive models, and we state our convergence results for this estimator in Section 3.2. We illustrate its convergence rates for various concrete instances of kernel classes. In Section 3.3, we provide a detailed comparison between our results to past and concurrent work, including discussion of the effect of global boundedness conditions on optimal rates. Finally, in Section 3.4, we state minimax lower bounds on the $L^2(\mathbb{P})$ error, which establish the optimality of our procedure.

3.1 A regularized M-estimator for sparse additive models

For any function of the form $f = \sum_{j=1}^{d} f_j$, the $(L^2(\mathbb{Q}_n), 1)$ and $(\mathcal{H}, 1)$-norms are given by

$$\|f\|_{n,1} := \sum_{j=1}^{d} \|f_j\|_n, \quad \text{and} \quad \|f\|_{\mathcal{H},1} := \sum_{j=1}^{d} \|f_j\|_{\mathcal{H}}, \tag{7}$$

respectively. Using this notation, we define the cost functional

$$\mathcal{L}(f) = \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda_n \|f\|_{n,1} + \rho_n \|f\|_{\mathcal{H},1}. \tag{8}$$

The cost functional $\mathcal{L}(f)$ is least-squares loss with a sparsity penalty $\|f\|_{n,1}$ and a smoothness penalty $\|f\|_{\mathcal{H},1}$. Here (λ_n, ρ_n) are a pair of positive regularization parameters whose choice will be specified by our theory. Given this cost functional, we then consider the M-estimator

$$\widehat{f} \in \arg\min_{f} \mathcal{L}(f) \quad \text{subject to } f = \textstyle\sum_{j=1}^{d} f_j \text{ and } \|f_j\|_{\mathcal{H}} \leq 1 \text{ for all } j = 1, 2, \ldots, d. \tag{9}$$

As stated, the problem (9) is infinite-dimensional in nature, since it involves optimization over Hilbert spaces. However, an attractive feature of this M-estimator is that, as a straightforward consequence of the representer theorem [16, 30], it can be reduced to an equivalent convex program in $\mathbb{R}^n \times \mathbb{R}^d$. In particular, for each $j = 1, 2, \ldots, d$, let $\mathbb{K}^j$ denote the kernel function for co-ordinate j. Using the notation $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ for the i^{th} sample, we define the collection of empirical kernel matrices $K^j \in \mathbb{R}^{n \times n}$, with entries $K^j_{i\ell} = \mathbb{K}^j(x_{ij}, x_{\ell j})$. By the representer theorem, any solution $\widehat{f}$ to the variational problem (9) can be written in the form

$$\widehat{f}(z_1, \ldots, z_d) = \sum_{i=1}^{n} \sum_{j=1}^{d} \widehat{\alpha}_{ij} \mathbb{K}^j(z_j, x_{ij}),$$

for a collection of weights $\big\{\widehat{\alpha}_j \in \mathbb{R}^n, \ j = 1, \ldots, d\big\}$. The optimal weights are obtained by solving the convex program

$$(\widehat{\alpha}_1, \ldots, \widehat{\alpha}_d) = \arg \min_{\substack{\alpha_j \in \mathbb{R}^n \\ \alpha_j^T K^j \alpha_j \leq 1}} \left\{ \frac{1}{2n} \|y - \sum_{j=1}^{d} K^j \alpha_j\|_2^2 + \lambda_n \sum_{j=1}^{d} \sqrt{\frac{1}{n} \|K^j \alpha_j\|_2^2} + \rho_n \sum_{j=1}^{d} \sqrt{\alpha_j^T K^j \alpha_j} \right\}. \tag{10}$$

This problem is a second-order cone program (SOCP), and there are various algorithms for solving it to arbitrary accuracy in time polynomial in (n, d), among them interior point methods (e.g., see §11 in the book [6]).

Various combinations of sparsity and smoothness penalties—all slightly different than the approach proposed here—have been used in in past work on sparse additive models. For instance, the method of Ravikumar et. al [28] is based least-squares loss regularized with single sparsity constraint, and separate smoothness constraints for each univariate function. They solve the resulting optimization problem using a back-fitting procedure. Koltchinskii and Yuan [17] develop a method

based on least-squares loss combined with a single penalty term $\sum_{j=1}^{d} \|f_j\|_{\mathcal{H}}$. Their method also leads to an SOCP if $\mathcal{H}$ is a reproducing kernel Hilbert space, but differs from the program (10) in lacking the additional sparsity penalties. Meier et. al [23] analyzed least-squares regularized with a penalty term of the form $\sum_{j=1}^{d} \sqrt{\lambda_1 \|f_j\|_n^2 + \lambda_2 \|f_j\|_{\mathcal{H}}^2}$, where λ_1 and λ_2 are a pair of regularization parameters. In their method, λ_1 controls the sparsity while λ_2 controls the smoothness. If $\mathcal{H}$ is an RKHS, the method in Meier et. al [23] reduces to an ordinary group Lasso problem on a different set of variables, which is another type of SOCP.

3.2 Convergence rates

We now state a result that provides convergence rates for the estimator (9), or equivalently (10). To simplify presentation, we state our result in the special case that the univariate Hilbert space $\mathcal{H}_j, j = 1, \ldots, d$ are all identical, denoted by $\mathcal{H}$. The analysis and results extend in a straightforward manner to the general setting of distinct univariate Hilbert spaces, as we discuss following the statement of Theorem 1.

Let $\mu_1 \geq \mu_2 \geq \ldots \geq 0$ denote the non-negative eigenvalues of the kernel operator defining the univariate Hilbert space $\mathcal{H}$, as defined in equation (3), and define the function

$$\mathcal{R}_n(t) := \frac{1}{\sqrt{n}} \Big[\sum_{\ell=1}^{\infty} \min\{t^2, \mu_\ell\} \Big]^{1/2}. \tag{11}$$

For a constant $\kappa_0 > 0$ to be chosen, let $\nu_n > 0$ be the smallest positive solution to the inequality

$$\nu_n^2 \geq \kappa_0 \, \mathcal{R}_n(\nu_n). \tag{12}$$

We refer to ν_n as the *critical univariate rate*, as it is the minimax-optimal rate for $L^2(\mathbb{P})$-estimation of a single univariate function in the Hilbert space $\mathcal{H}$ (e.g., [24, 32]). This quantity will be referred to throughout the remainder of the paper.

Our choices of regularization parameters are specified in terms of the quantity

$$\gamma_n := \kappa_1 \, \max\Big\{\nu_n, \sqrt{\frac{\log d}{n}}\Big\}, \tag{13}$$

where $\kappa_1 > 0$ is a sufficiently large constant, independent of the sample size and function classes. We assume that each function within the unit ball of the univariate Hilbert space is bounded—that is, for each $j = 1, \ldots, d$

$$\|f_j\|_\infty \leq 1 \qquad \text{for all } \|f_j\|_{\mathcal{H}} \leq 1. \tag{14}$$

This condition is fairly mild, and is implied by having a bounded univariate kernel function, for instance. These types of boundedness condition are quite standard for proving upper bounds on rates of convergence for non-parametric least squares in the univariate case $d = 1$ (see e.g. [31, 32]). However, note that we do not assume that the functions $f = \sum_{j \in S} f_j$ in $\mathcal{F}$ are uniformly bounded independently of (d, s).

The following result applies to any class $\mathcal{F}_{d,s,\mathcal{H}}$ of sparse additive models based on the univariate Hilbert space satisfying condition (14), and to the estimator (9) based on n i.i.d. samples $(x_i, y_i)_{i=1}^n$ from the observation model (6).

Theorem 1. *Let $\widehat{f}$ be any minimizer of the convex program (9) with regularization parameters $\lambda_n = c_3\gamma_n$ and $\rho_n = c_4\gamma_n^2$ for sufficiently large constants c_3 and c_4. Then provided that $n\gamma_n^2 = \Omega(\log(1/\gamma_n))$, there are universal constants (C, c_1, c_2) such that*

$$\mathbb{P}\Big[\|\widehat{f} - f^*\|_n^2 \geq C\big\{\frac{s\log d}{n} + s\nu_n^2\big\}\Big] \leq c_1\exp(-c_2 n\gamma_n^2). \tag{15}$$

Remarks: (a) The technical condition $n\gamma_n^2 = \Omega(\log(1/\gamma_n))$ is quite mild, and satisfied in most cases of interest, among them the kernels considered below in Corollaries 1 and 2.

(b) Although Theorem 1 is stated for the empirical $L^2(\mathbb{P}_n)$ error, the same bound holds for the error $\|\Pi_{\mathcal{F}}(\widehat{f}) - f^*\|_2$, where $\Pi_{\mathcal{F}}(\widehat{f})$ is the projection of $\widehat{f}$ onto the class $\mathcal{F}$ under the $L^2(\mathbb{P}_n)$-norm. Although we suspect that the error $\|\widehat{f} - f^*\|_2$ satisfies the same bound (15), our current techniques only allow us to control the projected function $\Pi_{\mathcal{F}}(\widehat{f})$. If we imposed a global boundedness condition, then it would follow that $\|\widehat{f} - f^*\|_2$ has the same scaling as $\|\widehat{f} - f^*\|_n$ under the given conditions. It remains an open question if one can directly establish such a bound without a global boundedness condition.

(c) For clarity, we have stated our result in the case where the univariate Hilbert space $\mathcal{H}$ is identical across all co-ordinates. However, our proof extends with only notational changes to the general setting, in which each co-ordinate j is endowed with a (possibly distinct) Hilbert space $\mathcal{H}_j$. In this case, the M-estimator returns a function $\widehat{f}$ such that (with high probability)

$$\|\widehat{f} - f^*\|_n^2 \quad \frac{s\log d}{n} + \sum_{j\in S}\nu_{n,j}^2,$$

where $\nu_{n,j}$ is the critical univariate rate associated with the Hilbert space $\mathcal{H}_j$, and S is the subset on which f^* is supported.

(d) As described in the introduction, the rate $\frac{s\log d}{n} + s\nu_n^2$ may be interpreted as the sum of a subset selection term ($\frac{s\log d}{n}$) and an s-dimensional estimation term ($s\nu_n^2$). Note that the subset selection term ($\frac{s\log d}{n}$) is independent of the choice of Hilbert space $\mathcal{H}$ whereas the s-dimensional estimation term is independent of the ambient dimension d. Depending on the scaling of the triple (n, d, s) and the smoothness of the univariate RKHS $\mathcal{H}$, either the subset selection term or function estimation term may dominate. In general, if $\frac{\log d}{n} = o(\nu_n^2)$, the s-dimensional estimation term dominates, and vice versa otherwise. At the boundary, the scalings of the two terms are equivalent.

Theorem 1 has a number of corollaries, obtained by specifying particular choices of kernels. First, we discuss m-rank operators, meaning that the kernel function $\mathbb{K}$ can be expanded in terms of m eigenfunctions. This class includes linear functions, polynomial functions, as well as any function class based on finite dictionary expansions.

Corollary 1. *Under the same conditions as Theorem 1, consider an univariate kernel with finite rank m. Then any solution $\widehat{f}$ to the problem (9) satisfies*

$$\mathbb{P}\Big[\max\big\{\|\widehat{f} - f^*\|_n^2, \|\Pi_{\mathcal{F}}(\widehat{f}) - f^*\|_2^2\big\} \geq C\big\{\frac{s\log d}{n} + s\frac{m}{n}\big\}\Big] \leq c_1\exp\big(-c_2(m + \log d)\big). \tag{16}$$

Proof. It suffices to show that the critical univariate rate (12) satisfies the scaling $\nu_n^2 = \mathcal{O}(m/n)$. For a finite-rank kernel and any $t > 0$, we have

$$\mathcal{R}_n(t) = \frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^{m} \min\{t^2, \mu_j\}} \; \leq \; t \sqrt{\frac{m}{n}},$$

from which the claim follows by the definition (12). □

Next, we present a result for the RKHS's with infinitely many eigenvalues, but whose eigenvalues decay at a rate $\mu_\ell \asymp (1/\ell)^{2\alpha}$ for some parameter $\alpha > 1/2$. Among other examples, this type of scaling covers the case of Sobolev spaces, say consisting of functions with α derivatives (e.g., [5, 13]).

Corollary 2. *Under the same conditions as Theorem 1, consider an univariate kernel with eigenvalue decay $\mu_\ell \asymp (1/\ell)^{2\alpha}$ for some $\alpha > 1/2$. Then the kernel estimator defined in (9) satisfies*

$$\mathbb{P}\Big[\max\big\{ \|\widehat{f} - f^*\|_n^2, \|\Pi_{\mathcal{F}}(\widehat{f}) - f^*\|_2^2 \big\} \geq C\big\{ \frac{s \log d}{n} + s\big(\frac{1}{n}\big)^{\frac{2\alpha}{2\alpha+1}} \big\} \Big] \leq c_1 \exp\big(- c_2 (n^{\frac{1}{2\alpha+1}} + \log d) \big). \tag{17}$$

Proof. As in the previous corollary, we need to compute the critical univariate rate ν_n. Given the assumption of polynomial eigenvalue decay, a truncation argument shows that $\mathcal{R}_n(t) = \mathcal{O}\big(\frac{t^{1-\frac{1}{2\alpha}}}{\sqrt{n}}\big)$. Consequently, the critical univariate rate (12) satisfies the scaling $\nu_n^2 \asymp \nu_n^{1-\frac{1}{2\alpha}}/\sqrt{n}$, or equivalently, $\nu_n^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}}$. □

3.3 Comparison with other work

It is interesting to compare these convergence rates in $L^2(\mathbb{P}_n)$ error with those established in past work [17, 23, 28]. Ravikumar et. al [28] show that any solution to their back-fitting method is consistent in terms of mean-squared error risk (see Theorem 3 in their paper). However, their analysis does not appear to track s explicitly, and assumes that d is sufficiently large to ensure that the subset selection term dominates, so the result is not directly comparable. The method of Koltchinskii and Yuan [17] is based regularizing the least-squares loss with the $(\mathcal{H}, 1)$-norm penalty—that is, $\sum_{j=1}^d \|f_j\|_{\mathcal{H}}$; Theorem 2 in their paper presents a rate that captures the decomposition into two terms, which can be interpreted as related to subset selection and s-dimensional estimation term. In quantitative terms, however, their rates are looser than those given here; in particular, their bound includes a term of the order $\frac{s^3 \log d}{n}$, which is larger than the bound in Theorem 1. For their algorithm, Meier et al. [23] establish a convergence rate of the form $\mathcal{O}(s(\frac{\log d}{n})^{\frac{2\alpha}{2\alpha+1}})$ in the case of α-smooth Sobolev spaces (see Theorem 1 in their paper). This result is sub-optimal compared to the optimal rate proven in Theorem 2(b) in regimes when d is large.[1] In all of the above-mentioned methods, it is unclear whether or not sharper analysis would yield better rates.

Finally, as discussed previously in the introduction, the concurrent work of Koltchinskii and Yuan [18] analyzes a method that is essentially the same as our M-estimator (9). In terms of rates

[1] More precisely, we either have $\frac{\log d}{n} < (\frac{\log d}{n})^{\frac{2\alpha}{2\alpha+1}}$, when subset selection term dominates, or $(\frac{1}{n})^{\frac{2\alpha}{2\alpha+1}} < (\frac{\log d}{n})^{\frac{2\alpha}{2\alpha+1}}$, when the s-dimensional estimation term dominates.

obtained, they establish a convergence rate based on two terms as in Theorem 1, but with a pre-factor that depends on the global bound $C(\mathcal{F}_{d,s}) = \sup_{f \in \mathcal{F}_{d,s}} \|f\|_\infty$. (Recall that functions $f \in \mathcal{F}_{d,s}$ consist of sums of the form $f = \sum_{j \in S} f_j$, where S has cardinality s.) In contrast, our pre-factor contains no dependence on this global quantity. Thus, if one assumes that $C(\mathcal{F}_{d,s}) = \mathcal{O}(1)$ even as (d, s) scale, then the rates obtained are the same up to constant factors. However, making such an assumption in the high-dimensional setting can be quite restrictive. Indeed, as shown by the following example, it can lead to quite "small" function classes $\mathcal{F}_{d,s}$ for which much faster rates can be achieved using different methods.

Example 1 (Restrictiveness of assuming global boundedness)**.** Suppose that each covariate x_j is uniform on $[-1, +1]$, and consider the class of univariate linear functions

$$\mathcal{H} := \{g_\alpha : \mathbb{R} \to \mathbb{R} \mid \alpha \in \mathbb{R}\}, \quad \text{where } g_\alpha(x_j) = \alpha\, x_j.$$

Thus, our function class $\mathcal{F} = \mathcal{F}_{d,s}$ consists of sparse linear functions of the form

$$f_\beta(x) = \sum_{j \in S} g_{\beta_j}(x) \;=\; \sum_{j \in S} \beta_j x_j.$$

Since $\|g_{\beta_j}\|_\infty = |\beta_j|$, boundedness of the univariate classes amounts to the requirement $|\beta_j| \leq 1$. Moreover, for any function $f_\beta \in \mathcal{F}$, note that we have $\|f_\beta\|_\infty = \|\beta\|_1$. Consequently, imposing the global boundedness condition $C(\mathcal{F}) = \sup_{f_\beta \in \mathcal{F}} \|f_\beta\|_\infty \leq R$ is equivalent to the constraint $\|\beta\|_1 \leq R$, so that the problem reduces to ordinary linear regression over the ℓ_1-ball $\mathbb{B}_1(R) = \{\beta \in \mathbb{R}^d \mid \|\beta\|_1 \leq R\}$. For this problem, it is known [8, 27] that the Lasso will produce an estimate such that

$$\|\widehat{\beta} - \beta^*\|_n^2 \;\; R \sqrt{\frac{\log d}{n}} \tag{18}$$

with high probability.[2] This rate is independent of s because the global boundedness condition restricts us to a ℓ_1-ball with constant radius R; indeed, the rate (18) is minimax-optimal over the set $\mathbb{B}_1(R)$ (e.g., see the paper [27]). In contrast, the error bound

$$\|\widehat{\beta} - \beta^*\|_n^2 \; \frac{s}{n} + \frac{s \log d}{n} \tag{19}$$

does depend on s and so can be substantially weaker, depending on the choice of s. For example, taking $s = \lceil \sqrt{d} \rceil$, the optimal rate (18) scales logarithmically in d whereas the scaling of the sub-optimal rate (19) is exponentially larger. This construction shows that the rates derived under global boundedness conditions are not minimax-optimal in general.

Returning to the setting of a general RKHS $\mathcal{H}(S)$, a global boundedness condition imposes an upper bound on the Hilbert norm radius of functions in $\mathcal{F}_{s,d} = \cup_{|S|=s} \mathcal{H}(S)$. Indeed, for a given subset S, let $\rho > 0$ be the largest radius such that $\{\|f\|_{\mathcal{H}(S)} \leq \rho\} \subseteq \mathcal{F}$. Then for any $x \in \mathcal{X}$, we

[2] Here we use to denote inequality up to constant factors depending on variances of the design and noise. This is the optimal rate for regression over ℓ_1-balls, as opposed to ℓ_0-balls.

have

$$\begin{aligned}\sup_{f\in\mathcal{F}}|f(x)| &\geq \sup_{\|f\|_{\mathcal{H}(S)}\leq\rho}|f(x)| \\ &= \sup_{\|f\|_{\mathcal{H}(S)}\leq\rho}|\langle f,\, \widetilde{\mathbb{K}}(\cdot,x)\rangle_{\mathcal{H}(S)}| \\ &= \rho\,\|\widetilde{\mathbb{K}}(\cdot,x)\|_{\mathcal{H}(S)} \\ &= \rho\sqrt{\widetilde{\mathbb{K}}(x,x)},\end{aligned}$$

where $\widetilde{\mathbb{K}}$ denotes the kernel associated with $\mathcal{H}(S)$. By definition of the Hilbert space $\mathcal{H}(S)$, we have $\widetilde{\mathbb{K}}(x,x) = \sum_{j\in S}\mathbb{K}(x_j,x_j)$, where $\mathbb{K}(x_j,x_j)$ is the univariate kernel over co-ordinate j. Consequently, we have the lower bound

$$\begin{aligned}\rho\sup_{x\in\mathcal{X}^{|S|}}\|\widetilde{\mathbb{K}}(\cdot,x)\|_{\mathcal{H}(S)} &= \rho\sup_{x\in\mathcal{X}^{|S|}}\sqrt{\sum_{j\in S}\mathbb{K}(x_j,x_j)} \\ &\geq \rho\,\sqrt{s}\sup_{x_1\in\mathcal{X}}\sqrt{\mathbb{K}(x_1,x_1)},\end{aligned}$$

showing that we require the bound $\rho = \mathcal{O}(1/\sqrt{s})$ in order to ensure $C(\mathcal{F}_{s,d}) = \mathcal{O}(1)$.

3.4 Minimax lower bounds

In this section, we provide minimax lower bounds in $L^2(\mathbb{P})$ error so as to complement the achievability results derived in Theorem 1. Given the function class $\mathcal{F}$, the minimax $L^2(\mathbb{P})$-error is given by

$$\mathfrak{M}_{\mathbb{P}}(\mathcal{F}) := \inf_{\widehat{f}_n}\sup_{f^*\in\mathcal{F}}\|\widehat{f}_n - f^*\|_2^2, \tag{20}$$

where the infimum is taken over all measureable functions of n samples $\{(y_i,x_i)\}_{i=1}^n$. As defined, this minimax error is a random variable, and our goal is to obtain a lower bound in probability.

Central to our proof of the lower bounds is the metric entropy structure of the univariate reproducing kernel Hilbert spaces. More precisely, our lower bounds depend on the *packing entropy,* defined as follows. Let $(\mathcal{G},\rho)$ be a totally bounded metric space, consisting of a set $\mathcal{G}$ and a metric $\rho : \mathcal{G}\times\mathcal{G}\to\mathbb{R}_+$. An ϵ-packing of $\mathcal{G}$ is a collection $\{f^1,\ldots,f^M\}\subset\mathcal{G}$ such that $\rho(f^i,f^j)\geq\epsilon$ for all $i\neq j$. The ϵ-packing number $M(\epsilon;\mathcal{G},\rho)$ is the cardinality of the largest ϵ-packing. The packing entropy is the simply the logarithm of the packing number, namely the quantity $\log M(\epsilon;\mathcal{G},\rho)$, to which we also refer as the metric entropy.

With this set-up, we derive explicit minimax lower bounds for two different scalings of the univariate metric entropy.

Logarithmic metric entropy: There exists some $m > 0$ such that

$$\log M(\epsilon;\mathbb{B}_{\mathcal{H}}(1),\|\cdot\|_2) \simeq m\,\log(1/\epsilon) \qquad \text{for all } \epsilon\in(0,1). \tag{21}$$

Function classes with metric entropy of this type include linear functions (for which $m = k$), univariate polynomials of degree k (for which $m = k+1$), and more generally, any function space

with finite VC-dimension [33]. This type of scaling also holds for any RKHS based on a kernel with rank m (e.g., see [10]), and these finite-rank kernels include both linear and polynomial functions as special cases.

Polynomial metric entropy There exists some $\alpha > 0$ such that

$$\log M(\epsilon; \mathbb{B}_{\mathcal{H}}(1), \|\cdot\|_2) \asymp (1/\epsilon)^{1/\alpha} \qquad \text{for all } \epsilon \in (0,1). \tag{22}$$

Various types of Sobolev/Besov classes exhibit this type of metric entropy decay [5, 13]. In fact, any RKHS in which the kernel eigenvalues decay at a rate $j^{-2\alpha}$ have a metric entropy with this scaling [9, 10].

We are now equipped to state our lower bounds on the minimax risk (20):

Theorem 2. *Given n i.i.d. samples from the sparse additive model (6) with sparsity $s \leq d/4$, there is an universal constant $C > 0$ such that:*

(a) For a univariate class $\mathcal{H}$ with logarithmic metric entropy (21) indexed by parameter m, we have

$$\mathfrak{M}_{\mathbb{P}}(\mathcal{F}) \;\geq\; C\left\{\frac{s\log(d/s)}{n} + s\,\frac{m}{n}\right\} \tag{23}$$

with probability greater than $1/2$.

(b) For a univariate class $\mathcal{H}$ with polynomial metric entropy (22) indexed by α, we have

$$\mathfrak{M}_{\mathbb{P}}(\mathcal{F}) \;\geq\; C\left\{\frac{s\log(d/s)}{n} + s\,\left(\frac{1}{n}\right)^{\frac{2\alpha}{2\alpha+1}}\right\} \tag{24}$$

with probability greater than $1/2$.

The choice of stating bounds that hold with probability $1/2$ is simply a convention often used in information-theoretic approaches (see, for instance, the papers [14, 35, 36]). We note that analogous lower bounds can established with probabilities arbitrarily close to one, albeit at the expense of worse constants. The most important consequence of Theorem 2 is in establishing the minimax-optimality of the results given in Corollary 1 and 2; in particular, in the regime sub-linear sparsity (i.e., for which $\log d = \mathcal{O}(\log(d/s))$), the combination of Theorem 2 with these corollaries identifies the minimax rates up to constant factors.

4 Proofs

In this section, we provide the proofs of our main results, namely Theorems 1 and 2. For clarity in presentation, we split the proofs up into a series of lemmas, with the bulk of the more technical proofs deferred to the appendices. This splitting allows our presentation in the main text to be relatively streamlined.

4.1 Proof of Theorem 1

At a high-level, Theorem 1 is based on an appropriate adaptation to the non-parametric setting of various techniques that have been developed for bounding the error in sparse linear regression (e.g., [4, 26]). In contrast to the parametric setting (where classical tail bounds are sufficient), controlling the error terms in this analysis requires more advanced techniques from empirical process theory. In particular, we make use of concentration theorems for Gaussian and empirical processes (e.g., [19, 22]) as well as results on the Rademacher complexity of kernel classes [24]. At a high-level, the proof is based on four technical lemmas. First, Lemma 1 provides an upper bound on the Gaussian complexity of any function of the form $f = \sum_{j=1}^{d} f_j$ in terms of the norms $\|\cdot\|_{\mathcal{H},1}$ and $\|\cdot\|_{n,1}$ previously defined. Lemma 2 exploits the notion of decomposability [26], as applied to these norms, in order to show that the error function belongs to a particular cone-shaped set. Finally, Lemma 3 and 4 establish some relations between the that the $L^2(\mathbb{P})$ and $L^2(\mathbb{P}_n)$ norms of functions in the class $\mathcal{F}$. The latter lemma involves a truncation argument so as to avoid having to impose global bounds on the function class.

Throughout the proof, we use C and c_i, $i = 1, 2, 3, 4$ to denote universal constants, independent of (n, d, s). Note that the precise numerical values of these constants may change from line to line. We use $(\kappa_0, \kappa_1, \kappa_2, \kappa_3)$ to denote constants, independent of (n, d, s), but whose value is fixed throughout. To ease notation, we define

$$\delta_n^2 := \kappa_2 \big\{ \frac{s \log d}{n} + s\nu_n^2 \big\},$$

where the constant $\kappa_2 > 0$ is to be chosen. Recall the definitions of ν_n and γ_n from equations (12) and (13) respectively, and note that $\delta_n = \Theta(\sqrt{s}\gamma_n)$. For a subset $A \subseteq \{1, 2, \ldots, d\}$ and an additively decomposed function $f = \sum_{j=1}^{d} f_j$, we adopt the convenient notation

$$\|f_A\|_{n,1} := \sum_{j \in A} \|f_j\|_n, \quad \text{and} \quad \|f_A\|_{\mathcal{H},1} := \sum_{j \in A} \|f_j\|_{\mathcal{H}}. \tag{25}$$

4.1.1 Establishing a basic inequality

We begin by establishing a basic inequality on the error function $\widehat{\Delta} := \widehat{f} - f^*$. Since $\widehat{f}$ and f^* are, respectively, optimal and feasible for the problem (9), we are guaranteed that $\mathcal{L}(\widehat{f}) \leq \mathcal{L}(f^*)$, and hence that the error function $\widehat{\Delta}$ satisfies the bound

$$\frac{1}{2n} \sum_{i=1}^{n} (w_i - \widehat{\Delta}(x_i))^2 + \lambda_n \|\widehat{f}\|_{n,1} + \rho_n \|\widehat{f}\|_{\mathcal{H},1} \leq \frac{1}{2n} \sum_{i=1}^{n} w_i^2 + \lambda_n \|f^*\|_{n,1} + \rho_n \|f^*\|_{\mathcal{H},1}.$$

Some simple algebra yields the bound

$$\frac{1}{2} \|\widehat{\Delta}\|_n^2 \leq \big| \frac{1}{n} \sum_{i=1}^{n} w_i \widehat{\Delta}(x_i) \big| + \lambda_n \|\widehat{\Delta}\|_{n,1} + \rho_n \|\widehat{\Delta}\|_{\mathcal{H},1}, \tag{26}$$

which we refer to as our basic inequality [32].

4.1.2 Controlling the noise term

The following lemma provides control the term on the right-hand side of inequality (26) by simultaneously bounding the Gaussian complexity for univariate function $\widehat{\Delta}_j$ in terms of their $\|.\|_n$ and $\|.\|_{\mathcal{H}}$ norms. In particular, recalling that $\gamma_n = \kappa_1 \max\{\sqrt{\frac{\log d}{n}}, \nu_n\}$, we have the following lemma.

Lemma 1. *For a constant $\kappa_3 > 0$, define the event*

$$\mathcal{T}(\gamma_n) := \left\{ \forall \;\; j = 1, 2, \ldots, d, \; \left| \frac{1}{n} \sum_{i=1}^{n} w_i \widehat{\Delta}_j(x_{ij}) \right| \leq 2\,\kappa_3 \;\; \left\{ \gamma_n^2 \, \|\widehat{\Delta}_j\|_{\mathcal{H}} + \gamma_n \, \|\widehat{\Delta}_j\|_n \right\} \right\}. \tag{27}$$

Then under the condition $n\gamma_n^2 = \Omega(\log(1/\gamma_n))$, we have

$$\mathbb{P}(\mathcal{T}(\gamma_n)) \geq 1 - c_1 \exp(-c_2 n \gamma_n^2). \tag{28}$$

The proof of this lemma, provided in Appendix A, uses concentration of measure for Lipschitz functions over Gaussian random variables [19] combined with a peeling argument [1, 32]. The subset selection term ($\frac{s \log d}{n}$) in Theorem 1 arises from taking the maximum over all d components.

4.1.3 Exploiting decomposability

The remainder of our analysis involves conditioning on the event $\mathcal{T}(\gamma_n)$. Using Lemma 1, on the event $\mathcal{T}(\gamma_n)$ we have:

$$\frac{1}{2}\|\widehat{\Delta}\|_n^2 \;\; \leq \;\; 2\kappa_3\gamma_n^2\|\widehat{\Delta}\|_{n,1} + 2\kappa_3\gamma_n^2\|\widehat{\Delta}\|_{\mathcal{H},1} + \lambda_n\|\widehat{\Delta}\|_{n,1} + \rho_n\|\widehat{\Delta}\|_{\mathcal{H},1}.$$

Recalling that S denotes the true support of the unknown function f^*, note that we have $\|\widehat{\Delta}\|_{n,1} = \|\widehat{\Delta}_S\|_{n,1} + \|\widehat{\Delta}_{S^c}\|_{n,1}$, with a similar decomposition for $\|\widehat{\Delta}\|_{\mathcal{H},1}$. The next lemma shows that conditioned on $\mathcal{T}(\gamma_n)$, the quantities $\|\widehat{\Delta}\|_{\mathcal{H},1}$ and $\|\widehat{\Delta}\|_{n,1}$ are not significantly larger than the corresponding norms as applied to the function $\widehat{\Delta}_S$.

Lemma 2. *Conditioned on $\mathcal{T}(\gamma_n)$, and with the choices $\lambda_n \geq 4\kappa_3\gamma_n$ and $\rho_n \geq 4\kappa_3\gamma_n^2$, we have*

$$\lambda_n\|\widehat{\Delta}\|_{n,1} + \rho_n\|\widehat{\Delta}\|_{\mathcal{H},1} \leq 4\lambda_n\|\widehat{\Delta}_S\|_{n,1} + 4\rho_n\|\widehat{\Delta}_S\|_{\mathcal{H},1} \tag{29}$$

The proof of this lemma, provided in Appendix B, is based on the decomposability [26] of the $\|\cdot\|_{\mathcal{H},1}$ and $\|\cdot\|_{n,1}$ norms. This lemma allows us to exploit the sparsity assumption, since in conjunction with Lemma 1, we have now bounded the right-hand side of the basic inequality (26) in terms involving only $\widehat{\Delta}_S$. In particular, still conditioning on $\mathcal{T}(\gamma_n)$ and applying Lemma 2, we obtain

$$\begin{aligned}\|\widehat{\Delta}\|_n^2 &\leq C\left\{\gamma_n\|\widehat{\Delta}_S\|_{n,1} + \gamma_n^2\|\widehat{\Delta}_S\|_{\mathcal{H},1} + \lambda_n\|\widehat{\Delta}_S\|_{n,1} + \rho_n\|\widehat{\Delta}_S\|_{\mathcal{H},1}\right\} \\ &\leq C\left\{\gamma_n\|\widehat{\Delta}_S\|_{n,1} + \gamma_n^2\|\widehat{\Delta}_S\|_{\mathcal{H},1}\right\},\end{aligned}$$

where[3] we have recalled our choices $\lambda_n = \Theta(\gamma_n)$ and $\rho_n = \Theta(\gamma_n^2)$. Finally, since both $\widehat{f}_j$ and f_j^* belong to $\mathbb{B}_{\mathcal{H}}(1)$, we have

$$\|\widehat{\Delta}_j\|_{\mathcal{H}} \leq \|\widehat{f}_j\|_{\mathcal{H}} + \|f_j^*\|_{\mathcal{H}} \leq 2,$$

which implies that $\|\widehat{\Delta}_S\|_{\mathcal{H},1} \leq 2s$, and hence

$$\|\widehat{\Delta}\|_n^2 \leq C\left\{\gamma_n\|\widehat{\Delta}_S\|_{n,1} + s\gamma_n^2\right\}. \tag{30}$$

[3] In this step and elsewhere, the reader should be reminded of our convention that the numerical value of C can change from line to line.

4.1.4 Relating the $L^2(\mathbb{P}_n)$ and $L^2(\mathbb{P})$ norms

It remains to control the term $\|\widehat{\Delta}_S\|_{n,1} = \sum_{j \in S} \|\widehat{\Delta}_j\|_n$. Ideally, we would like to upper bound it by $\sqrt{s}\|\widehat{\Delta}_S\|_n$. Such an upper bound would follow immediately if it were phrased in terms of the $\|\cdot\|_2$ rather than the $\|\cdot\|_n$ norm, but there are additional cross-terms with the empirical norm. Accordingly, we make use of two lemmas that relate the $\|\cdot\|_n$ norm and the population $\|\cdot\|_2$ norms for functions in $\mathcal{F}$.

In the statements of these results, we adopt the notation g and g_j (as opposed to f and f_j) to be clear our results apply to any $g \in \mathcal{F}$. We first provide an upper bound on the empirical norm $\|g_j\|_n$ in terms of the associated $\|g_j\|_2$ norm, one that holds uniformly over all components $j = 1, 2, \ldots, d$.

Lemma 3. *For a universal constant C and $j = 1, 2, \ldots, d$, consider the events*

$$\mathcal{A}_j(\gamma_n) := \big\{\|g_j\|_n \leq 4\|g_j\|_2 + C\gamma_n \quad \text{for all } g_j \in \mathbb{B}_{\mathcal{H}}(2)\big\}, \tag{31}$$

as well as $\mathcal{A}(\gamma_n) = \cap_{j=1}^d \mathcal{A}_j(\gamma_n)$. If the univariate Hilbert space $\mathcal{H}$ satisfies condition (14), then there are universal constants (c_1, c_2) such that

$$\mathbb{P}\big[\mathcal{A}(\gamma_n)\big] \geq 1 - c_1 \exp(-c_2 n\gamma_n^2).$$

We now define the function class $2\mathcal{F} := \{f + f' \mid f, f' \in \mathcal{F}\}$. Our second lemma guarantees that the empirical norm $\|\cdot\|_n$ of any function in $2\mathcal{F}$ is uniformly lower bounded by the norm $\|\cdot\|_2$.

Lemma 4. *Define the event*

$$\mathcal{B}(\delta_n) := \bigg\{\|g\|_n^2 \geq \|g\|_2^2/4 \quad \text{for all } g \in 2\mathcal{F} \text{ with } \|g\|_2 \geq \delta_n\bigg\}. \tag{32}$$

If the underlying univariate Hilbert space $\mathcal{H}$ satisfies condition (14), then there are universal constants (c_1, c_2) such that

$$\mathbb{P}[\mathcal{B}(\delta_n)] \geq 1 - c_1 \exp(-c_2 n\delta_n^2).$$

Lemmas 3 and 4 are proved in Appendices F and D, respectively. Note that while both results require bounds on the univariate function classes (recall condition (14)), they do not require global boundedness assumptions—that is, on quantities of the form $\|\sum_{j \in S} g_j\|_\infty$. Typically, we expect that the $\|\cdot\|_\infty$-norms of functions $g \in \mathcal{F}$ scale with s.

4.1.5 Completing the proof

Using Lemmas 3 and 4, we can complete the proof of Theorem 1. For the remainder of the proof, let us condition on the events $\mathcal{A}(\gamma_n)$ and $\mathcal{B}(\delta_n)$. Conditioning on the event $\mathcal{A}(\gamma_n)$, we have

$$\|\widehat{\Delta}_S\|_{n,1} = \sum_{j \in S} \|\widehat{\Delta}_j\|_n \leq 4 \sum_{j \in S} \|\widehat{\Delta}_j\|_2 + Cs\gamma_n \;\leq\; 4\sqrt{s}\|\widehat{\Delta}_S\|_2 + Cs\gamma_n. \tag{33}$$

Our next step is to upper bound $\|\widehat{\Delta}_S\|_2$ in terms of $\|\widehat{\Delta}_S\|_n$ and $s\gamma_n$. We split our analysis into two cases.

Case 1: If $\|\widehat{\Delta}_S\|_2 < \delta_n = \Theta(\sqrt{s}\gamma_n)$, then combined with the bound (33), we conclude that

$$\|\widehat{\Delta}_S\|_{n,1} \leq Cs\gamma_n. \tag{34}$$

Case 2: Otherwise, we have $\|\widehat{\Delta}_S\|_2 \geq \delta_n$. Note that the function $\widehat{\Delta}_S = \sum_{j\in S}\widehat{\Delta}_j$ belongs to the class $2\mathcal{F}$, so that it is covered by the event $\mathcal{B}(\delta_n)$. In particular, conditioned on the event $\mathcal{B}(\delta_n)$, we have $\|\widehat{\Delta}_S\|_2 \leq 2\|\widehat{\Delta}_S\|_n$. Combined with the bound (33), we conclude that

$$\|\widehat{\Delta}_S\|_{n,1} \leq C\{\sqrt{s}\|\widehat{\Delta}_S\|_n + s\gamma_n\}. \tag{35}$$

Note that (disregarding constants) the bound (34) is at least as good as the bound (35); therefore, in either case, a bound of the form (35) holds. Substituting the inequality (35) in the bound (30) yields

$$\|\widehat{\Delta}\|_n^2 \leq C\left\{\gamma_n\|\widehat{\Delta}_S\|_{n,1} + s\gamma_n^2\right\} \;\leq\; C\{\sqrt{s}\gamma_n\,\|\widehat{\Delta}_S\|_n + s\gamma_n^2\}. \tag{36}$$

Since $\|\widehat{\Delta}_S\|_n \leq \|\widehat{\Delta}\|_n$, the bound (36) implies that $\|\widehat{\Delta}\|_n \leq C\,\sqrt{s}\gamma_n$. This bound is valid conditioned on the events $\mathcal{T}(\gamma_n)$, $\mathcal{A}(\gamma_n)$ and $\mathcal{B}(\delta_n)$. Using Lemmas 1, 3 and 4 in conjunction, we obtain

$$\mathbb{P}(\mathcal{T}(\gamma_n) \cap \mathcal{A}(\gamma_n) \cap \mathcal{B}(\delta_n)) \geq 1 - c_1\exp\big(-c_2 n\gamma_n^2\big),$$

thereby showing that $\|\widehat{f} - f^*\|_n \leq C\sqrt{s}\gamma_n$ with the claimed probability.

Finally, let us extend the result to the error $\|\Pi_{\mathcal{F}}(\widehat{f}) - f^*\|_2$, as mentioned in the remarks following Theorem 1. In order to do so, we exploit Lemma 4. Since the function $\Pi_{\mathcal{F}}(\widehat{f}) - f^*$ belongs to $2\mathcal{F}$, we may apply the lemma to it. We conclude that either $\|\Pi_{\mathcal{F}}(\widehat{f}) - f^*\|_2 \leq \delta_n$, in which case we are done, or that, with probability at least $1 - c_1\exp(-c_2 n\delta_n^2)$, we have

$$\begin{aligned}\|\Pi_{\mathcal{F}}(\widehat{f}) - f^*\|_2 &\leq 2\|\Pi_{\mathcal{F}}(\widehat{f}) - f^*\|_n \\ &\leq 2\big\{\|\Pi_{\mathcal{F}}(\widehat{f}) - \widehat{f}\|_n + \|\widehat{f} - f^*\|_n\big\}\end{aligned}$$

where the second step follows by triangle inequality. Now by definition of the projection, since $f^* \in \mathcal{F} \subset 2\mathcal{F}$, we must have $\|\widehat{f} - f^*\|_n \geq \|\widehat{f} - \Pi_{\mathcal{F}}(\widehat{f})\|_n$, from which we conclude that

$$\|\Pi_{\mathcal{F}}(\widehat{f}) - f^*\|_2 \leq 4\|\widehat{f} - f^*\|_n,$$

which completes the proof of Theorem 1.

4.2 Proof of Theorem 2

We now turn to the proof of the minimax lower bounds stated in Theorem 2. For both parts (a) and (b), the first step is to follow a standard reduction to testing (e.g., [14, 35, 36]) so as to obtain a lower bound on the minimax error $\mathfrak{M}_{\mathbb{P}}(\mathcal{F})$ in terms of the probability of error in a multi-way hypothesis testing. We then apply different forms of the Fano inequality [36, 35] in order to lower bound the probability of error in this testing problem. Obtaining useful bounds requires a precise characterization of the metric entropy structure of $\mathcal{F}_{d,s,\mathcal{H}}$, as stated in Lemma 5.

4.2.1 Reduction to testing

We begin with the reduction to a testing problem. Let $\{f^1, \ldots, f^N\}$ be a δ_n-packing of $\mathcal{F}$ in the $\|\cdot\|_2$-norm, and let Θ be a random variable uniformly distributed over the index set $[N] := \{1, 2, \ldots, N\}$. Note that we are using N as a shorthand for the packing number $M(\delta_n; \mathcal{F}, \|\cdot\|_2)$. A standard argument (e.g., [14, 35, 36]) then yields the lower bound

$$\inf_{\widehat{f}} \sup_{f^* \in \mathcal{F}} \mathbb{P}\big[\|\widehat{f} - f^*\|_2^2 \geq \delta_n^2/2\big] \geq \inf_{\widehat{\Theta}} \mathbb{P}[\widehat{\Theta} \neq \Theta], \tag{37}$$

where the infimum on the right-hand side is taken over all estimators $\widehat{\Theta}$ that are measurable functions of the data, and take values in the index set $[N]$.

Note that $\mathbb{P}[\widehat{\Theta} \neq \Theta]$ corresponds to the error probability in a multi-way hypothesis test, where the probability is taken over the random choice of Θ, the randomness of the design points $X_1^n := \{x_i\}_{i=1}^n$, and the randomness of the observations $Y_1^n := \{y_i\}_{i=1}^n$. Our initial analysis is performed conditionally on the design points, so that the only remaining randomness in the observations Y_1^n comes from the observation noise $\{w_i\}_{i=1}^n$. From Fano's inequality [11], for any estimator $\widehat{\Theta}$, we have $\mathbb{P}\big[\widehat{\Theta} \neq \Theta \mid X_1^n\big] \geq 1 - \frac{I_{X_1^n}(\Theta; Y_1^n) + \log 2}{\log N}$, where $I_{X_1^n}(\Theta; Y_1^n)$ denotes the mutual information between Θ and Y_1^n with X_1^n fixed. Taking expectations over X_1^n, we obtain the lower bound

$$\mathbb{P}\big[\widehat{\Theta} \neq \Theta\big] \geq 1 - \frac{\mathbb{E}_{X_1^n}\big[I_{X_1^n}(\Theta; Y_1^n)\big] + \log 2}{\log N}. \tag{38}$$

The remainder of the proof consists of constructing appropriate packing sets of $\mathcal{F}$, and obtaining good upper bounds on the mutual information term in the lower bound (38).

4.2.2 Constructing appropriate packings

We begin with results on packing numbers. Recall that $\log M(\delta; \mathcal{F}, \|\cdot\|_2)$ denotes the δ-packing entropy of $\mathcal{F}$ in the $\|\cdot\|_2$ norm.

Lemma 5. *(a) For all $\delta \in (0, 1)$ and $s \leq d/4$, we have*

$$\log M(\delta; \mathcal{F}, \|\cdot\|_2) = \mathcal{O}\big(s \ \log M(\frac{\delta}{\sqrt{s}}; \mathbb{B}_{\mathcal{H}}(1), \|\cdot\|_2) + s \log \frac{d}{s}\big). \tag{39}$$

(b) For a Hilbert class with logarithmic metric entropy (21) and such that $\|f\|_2 \leq \|f\|_{\mathcal{H}}$, there exists set $\{f^1, \ldots, f^M\}$ with $\log M \ \geq \ C\big\{s \log(d/s) + sm\big\}$, and

$$\delta \ \leq \ \|f^k - f^m\|_2 \ \leq \ 8\delta \qquad \text{for all } k \neq m \in \{1, 2, \ldots, M\}. \tag{40}$$

The proof, provided in Appendix E, is combinatorial in nature. We now turn to the proofs of parts (a) and (b) of Theorem 2.

4.2.3 Proof of Theorem 2(a)

In order to prove this claim, it remains to exploit Lemma 5 in an appropriate way, and to upper bound the resulting mutual information. For the latter step, we make use of the generalized Fano approach (e.g., [36]).

From Lemma 5, we can find a set $\{f^1, \ldots, f^M\}$ that is a δ-packing of $\mathcal{F}$ in ℓ_2-norm, and such that $\|f^k - f^\ell\|_2 \leq 8\delta$ for all $k, \ell \in [M]$. For $k = 1, \ldots, M$, let $\mathbb{Q}^k$ denote the conditional distribution of Y_1^n conditioned on X_1^n and the event $\{\Theta = k\}$, and let $D(\mathbb{Q}^k \,\|\, \mathbb{Q}^\ell)$ denote the Kullback-Leibler divergence. From the convexity of mutual information [11], we have the upper bound $I_{X_1^n}(\Theta; Y_1^n) \leq \frac{1}{\binom{M}{2}} \sum_{k,\ell=1}^M D(\mathbb{Q}^k \,\|\, \mathbb{Q}^\ell)$. Given our linear observation model (6), we have

$$D(\mathbb{Q}^k \,\|\, \mathbb{Q}^\ell) = \frac{1}{2\sigma^2} \sum_{i=1}^n \big(f^k(x_i) - f^\ell(x_i)\big)^2 \;=\; \frac{n\,\|f^k - f^\ell\|_n^2}{2},$$

and hence

$$\mathbb{E}_{X_1^n}\big[I_{X_1^n}(Y_1^n; \Theta)\big] \leq \frac{n}{2} \frac{1}{\binom{M}{2}} \sum_{k,\ell=1}^M \mathbb{E}_{X_1^n}[\|f^k - f^\ell\|_n^2] \;=\; \frac{n}{2} \frac{1}{\binom{M}{2}} \sum_{k,\ell=1}^M \|f^k - f^\ell\|_2^2.$$

Since our packing satisfies $\|f^k - f^\ell\|_2^2 \leq 64\delta^2$, we conclude that

$$\mathbb{E}_{X_1^n}\big[I_{X_1^n}(Y_1^n; \Theta)\big] \leq 32 n \delta^2.$$

From the Fano bound (38), for any $\delta > 0$ such that $\frac{32n\delta^2 + \log 2}{\log M} < \frac{1}{4}$, then we are guaranteed that $\mathbb{P}[\widehat{\Theta} \neq \Theta] \geq \frac{3}{4}$. From Lemma 5(b), our packing set satisfies $\log M \geq C\big\{sm + s\log(d/s)\big\}$, so that so that the choice $\delta^2 = C'\big\{\frac{sm}{n} + \frac{s\log(d/s)}{n}\big\}$, for a suitably small $C' > 0$, can be used to guarantee the error bound $\mathbb{P}[\widehat{\Theta} \neq \Theta] \geq \frac{3}{4}$.

4.2.4 Proof of Theorem 2(b)

In this case, we use an upper bounding technique due to Yang and Barron [35] in order to upper bound the mutual information. Although the argument is essentially the same, it does not follow verbatim from their claims—in particular, there are some slight differences due to our initial conditioning—so that we provide the details here. By definition of the mutual information, we have

$$I_{X_1^n}(\Theta; Y_1^n) = \frac{1}{M} \sum_{k=1}^M D(\mathbb{Q}^k \,\|\, \mathbb{P}_Y),$$

where $\mathbb{Q}^k$ denotes the conditional distribution of Y_1^n given $\Theta = k$ and still with X_1^n fixed, whereas $\mathbb{P}_Y$ denotes the marginal distribution of $\mathbb{P}_Y$. Now let $\{g^1, \ldots, g^N\}$ be an ϵ-cover of $\mathcal{F}$ in the $\|\cdot\|_2$ norm, for a tolerance ϵ to be chosen. As argued in Yang and Barron [35], we have

$$I_{X_1^n}(\Theta; Y_1^n) \;=\; \frac{1}{M} \sum_{j=1}^M D(\mathbb{Q}^j \,\|\, \mathbb{P}_Y) \leq D(\mathbb{Q}^k \,\|\, \frac{1}{N} \sum_{k=1}^N \mathbb{P}^k),$$

where $\mathbb{P}^\ell$ denotes the conditional distribution of Y_1^n given g^ℓ and X_1^n. For each ℓ, let us choose $\ell^*(k) \in \arg\min_{\ell=1,\ldots,N} \|g^\ell - f^k\|_2$. We then have the upper bound

$$I_{X_1^n}(\Theta; Y_1^n) \leq \frac{1}{M} \sum_{k=1}^M \big\{\log N + \frac{n}{2} \|g^{\ell^*(k)} - f^k\|_n^2\big\}.$$

Taking expectations over X_1^n, we obtain

$$\begin{aligned}\mathbb{E}_{X_1^n}[I_{X_1^n}(\Theta; Y_1^n)] &\leq \frac{1}{M}\sum_{k=1}^{M}\Big\{\log N + \frac{n}{2}\mathbb{E}_{X_1^n}[\|g^{\ell^*(k)} - f^k\|_n^2]\Big\} \\ &\leq \log N + \frac{n}{2}\,\epsilon^2,\end{aligned}$$

where the final inequality follows from the choice of our covering set.

From this point, we can follow the same steps as Yang and Barron [35]. The polynomial scaling (22) of the metric entropy guarantees that their conditions are satisfied, and we conclude that the minimax error is lower bounded any $\delta_n > 0$ such that

$$n\delta_n^2 \quad \log N(\delta_n; \mathcal{F}, \|\cdot\|_2).$$

From Lemma 5 and the assumed scaling (22), it is equivalent to solve the equation

$$n\delta_n^2 \quad s\log(d/s) + s(\sqrt{s}/\delta_n)^{1/\alpha},$$

from which some algebra yields $\delta_n^2 = C\big\{\frac{s\log(d/s)}{n} + s\big(\frac{1}{n}\big)^{\frac{2\alpha}{2\alpha+1}}\big\}$ as a suitable choice.

5 Discussion

In this paper, we have studied estimation in the class of sparse additive models defined by univariate reproducing kernel Hilbert spaces. In conjunction, our two main results provide a precise characterization of the minimax-optimal rates for estimating f^* in the $L^2(\mathbb{P})$-norm for various kernel classes. These classes include the case of finite-rank kernels (with logarithmic metric entropy), as well as kernels with polynomially decaying eigenvalues (and hence polynomial metric entropy). In order to establish achievable rates, we analyzed a simple M-estimator based on regularizing the least-squares loss with two kinds of ℓ_1-based norms, one defined by the univariate Hilbert norm and the other by the univariate empirical norm. On the other hand, we obtained our lower bounds by a combination of approximation-theoretic and information-theoretic techniques. An interesting feature of the minimax rates derived here is that they exhibit a natural decoupling into the complexities associated with two sub-problems. The first term corresponds to the difficulty of performing subset selection—that is, determining which s out of d co-ordinate functions are active. The second term corresponds to the difficulty of estimating a sum of s univariate functions, assuming that the correct co-ordinates are known.

There are a number of ways in which this work could be extended. For instance, although our analysis was based on assuming independence of the covariates x_j, $j = 1, 2, \ldots d$, it would be interesting to investigate the case when the random variables are endowed with some correlation structure. One might expect some changes in the optimal rates, particularly if many of the variables are strongly dependent. This work considered only the function class consisting of sums of univariate functions; a natural extension would be to consider nested non-parametric classes formed of sums over hierarchies of subsets of variables. Analysis in this case would require dealing with dependencies between the different functions.

Acknowledgements

This work was partially supported by NSF grants DMS-0605165 and DMS-0907632 to MJW and BY. In addition, BY was partially supported by the NSF grant SES-0835531 (CDI) and as well as a grant from the MSRA. MJW was also partially supported AFOSR Grant FA9550-09-1-0466. During this work, GR was financially supported by a Berkeley Graduate Fellowship.

A Proof of Lemma 1

Define the function

$$\widehat{\mathcal{R}}_{n,j}(r) \; := \mathbb{E}_w\big[\sup_{\substack{\|g_j\|_n \leq r \\ \|g_j\|_{\mathcal{H}} \leq 1}} \frac{1}{n} \sum_{i=1}^n w_i g(x_{ij}) \big],$$

and let $\widehat{\nu}_{n,j} > 0$ denote the smallest positive solution of the inequality $256\, r^2 \geq \widehat{\mathcal{R}}_{n,j}(r)$. The function $\widehat{\mathcal{R}}_{n,j}(r)$ defines the local Gaussian complexity of the kernel class in co-ordinate j. Using the techniques of Mendelson [24], it follows that there is an universal constant $c_0 > 0$ such that

$$\widehat{\mathcal{R}}_{n,j}(r) \leq \frac{c_0}{\sqrt{n}} \big[\sum_{j=1}^n \min\{\widehat{\mu}_j, r^2\} \big]^{1/2}, \tag{41}$$

where $\{\widehat{\mu}_\ell\}_{\ell=1}^n$ are the eigenvalues of the empirical kernel matrix. (The results of Mendelson are stated for the population Rademacher complexity, but a similar argument establishes the bound (41) for the empirical Gaussian complexity.)

Recall that the critical univariate rate ν_n is defined in terms of the closely related function $\mathcal{R}_n(r) := \frac{1}{\sqrt{n}} \big[\sum_{\ell=1}^\infty \min\{r^2, \mu_\ell\} \big]^{1/2}$, where $\{\mu_\ell\}_{\ell=1}^\infty$ are the eigenvalues of the (population) kernel operator. Define the event

$$\mathcal{D}(\gamma_n) := \big\{ \widehat{\nu}_{n,j} \leq \gamma_n, \quad \text{for all } j = 1, 2, \ldots, d \big\}, \tag{42}$$

where we recall that $\gamma_n := \kappa_1 \max\big\{\nu_n, \sqrt{\frac{\log d}{n}}\big\}$. It is a consequence of Lemma 6 in Appendix F that $\mathbb{P}[\mathcal{D}(\gamma_n)] \geq 1 - c_1 \exp(-c_2 n \gamma_n^2)$. Consequently, we proceed by conditioning on this event throughout the remainder of the proof.

In the remainder of the proof, our goal is to prove that

$$\big| \frac{1}{n} \sum_{i=1}^n w_i f_j(x_{ij}) \big| \leq C \, \big\{ \gamma_n^2 \, \|f_j\|_{\mathcal{H}} + \gamma_n \, \|f_j\|_n \big\} \qquad \text{for all } f_j \in \mathcal{H} \tag{43}$$

with probability greater than $1 - c_1 \exp(-c_2 n \gamma_n^2)$. By combining this result with our choice of γ_n and the union bound, the claimed bound on $\mathbb{P}[\mathcal{T}(\gamma_n)]$ then follows.

If $f_j = 0$, then the claim (43) is trivial. Otherwise, we write

$$\frac{1}{n} \sum_{i=1}^n w_i f_j(x_{ij}) = \|f_j\|_{\mathcal{H}} \, \frac{1}{n} \sum_{i=1}^n w_i g_j(x_{ij}), \qquad \text{where } g_j := f_j / \|f_j\|_{\mathcal{H}}.$$

Noting that $\|g_j\|_{\mathcal{H}} = 1$, we are led to study the random variable

$$Z_{n,j}(w; r_j) := \sup_{\substack{\|g_j\|_n \leq r_j \\ \|g_j\|_{\mathcal{H}} \leq 1}} \frac{1}{n} \sum_{i=1}^n w_i g_j(x_{ij}),$$

a quantity that satisfies $\widehat{\mathcal{R}}_{n,j}(r_j) = \mathbb{E}_w[Z_{n,j}(w; r_j)]$ by construction.

Our next step is to establish that for any fixed radius $r_j > 0$, we have tail bound

$$\mathbb{P}\Big[Z_{n,j}(w; r_j) \geq C\,\{\gamma_n^2 \; + \gamma_n \, r_j\}\Big] \leq c_1 \exp\big\{ - c_2 n \gamma_n^2 \big(1 + (\gamma_n / r_j)^2\big)\big\}. \tag{44}$$

We then use a peeling argument to extend the bound to a uniform one over the radius r_j.

Establishing the tail bound (44): Viewing $Z_{n,j}$ as a function of w, we first bound its Lipschitz constant. For any two vectors $w, w' \in \mathbb{R}^n$, we have

$$\begin{aligned} |Z_{n,j}(w; r_j) - Z_{n,j}(w'; r_j)| &\leq \frac{1}{n} \sup_{\|g_j\|_n \leq r_j} |\sum_{i=1}^n (w_i - w_i') g_j(x_{ij})| \\ &\leq \frac{r_j}{\sqrt{n}} \|w - w'\|_2 \end{aligned}$$

Therefore, by concentration of measure for Lipschitz functions of Gaussian variables [19], we have

$$\mathbb{P}\big[Z_{n,j}(w; r_j) \geq \mathbb{E}[Z_{n,j}(w; r_j)] + t\big] \leq 2 \exp\big(- n \frac{t^2}{2 r_j^2}\big). \tag{45}$$

Setting $t = \gamma_n(r_j + \gamma_n)$ yields an upper bound of the form of right-hand side of equation (44).

In order to complete the proof of the bound (44), we need to show

$$\mathbb{E}[Z_{n,j}(w; r_j)] \leq C\big\{\gamma_n^2 + \gamma_n \, r_j\big\} \qquad \text{for all } r_j > 0.$$

We do so by splitting into two cases.

Case 1: If $r_j \leq \widehat{\nu}_{n,j}$, then we have $\widehat{\mathcal{R}}_{n,j}(r_j) \; \leq \widehat{\mathcal{R}}_{n,j}(\widehat{\nu}_{n,j}) \; \leq 256\, \widehat{\nu}_{n,j}^2$ where the second inequality follows from our choice of $\widehat{\nu}_{n,j}$.

Case 2: Otherwise, if $r_j > \widehat{\nu}_{n,j}$, we have

$$\begin{aligned} \widehat{\mathcal{R}}_{n,j}(r) &= \frac{r}{\widehat{\nu}_{n,j}} \mathbb{E}_w\big[\sup_{\substack{\|g_j\|_n \leq \widehat{\nu}_{n,j} \\ \|g_j\|_{\mathcal{H}} \leq \frac{\widehat{\nu}_{n,j}}{r}}} \frac{1}{n} \sum_{i=1}^n w_i g(x_{ij})\big] \\ &\leq \frac{r}{\widehat{\nu}_{n,j}} \widehat{\mathcal{R}}_{n,j}(\widehat{\nu}_{n,j}) \\ &\leq 256\, r \widehat{\nu}_{n,j}, \end{aligned}$$

where the final line uses the fact that $\widehat{\mathcal{R}}_{n,j}(\widehat{\nu}_{n,j}) \leq 256\, \widehat{\nu}_{n,j}^2$. Combining yields the bound

$$\widehat{\mathcal{R}}_{n,j}(r) \leq C\big\{\widehat{\nu}_{n,j}^2 + r\, \widehat{\nu}_{n,j}\big\}.$$

Under the event $\mathcal{D}(\gamma_n)$ previously defined (42), we have $\widehat{\nu}_{n,j} \leq \gamma_n$, so that the proof of the claim (44) is complete.

Peeling argument: We now use the bound (44) to prove the bound (43), in particular via a "peeling" operation over all choices of $r_j = \|f_j\|_n/\|f_j\|_{\mathcal{H}}$. We first claim that it suffices to consider $r \in (0,1]$. In order to show that $r \leq 1$, it is equivalent to show that $\|g_j\|_n \leq 1$ for any $g_j \in \mathbb{B}_{\mathcal{H}}(1)$. Recall that we have assumed that $\|g_j\|_\infty \leq 1$ for all $g_j \in \mathbb{B}_{\mathcal{H}}(1)$. Consequently, whenever $g_j \in \mathbb{B}_{\mathcal{H}}(1)$, we have $\|g_j\|_n^2 = \frac{1}{n}\sum_{i=1}^n g_j^2(x_{ij}) \leq 1$, as required.

Now define the event

$$\mathcal{T}_j(\gamma_n) := \Big\{ \exists f_j \in \mathbb{B}_{\mathcal{H}}(1) \;\mid\; \big|\frac{1}{n}\sum_{i=1}^n w_i f_j(x_{ij})\big| > 2\,C\,\|f_j\|_{\mathcal{H}}\big\{\gamma_n^2 + \gamma_n \frac{\|f_j\|_n}{\|f_j\|_{\mathcal{H}}}\big\}\Big\}. \tag{46}$$

Note that we have the decomposition $\mathcal{T}_j(\gamma_n) = \mathcal{T}_j^A(\gamma_n) \cup \mathcal{T}_j^B(\gamma_n)$, where

$$\mathcal{T}_j^A(\gamma_n) := \mathcal{T}_j(\gamma_n) \cap \big\{\frac{\|f_j\|_n}{\|f_j\|_{\mathcal{H}}} \leq \gamma_n\big\}, \quad \text{and}$$
$$\mathcal{T}_j^B(\gamma_n) := \mathcal{T}_j(\gamma_n) \cap \big\{\frac{\|f_j\|_n}{\|f_j\|_{\mathcal{H}}} \in (\gamma_n, 1]\big\}.$$

It remains to obtain upper bounds on the probabilities of these two events.

Case A: For $m = 1, 2, 3, \ldots$, define the sets

$$S_m := \big\{\frac{\gamma_n}{2^m} \leq \frac{\|f_j\|_n}{\|f_j\|_{\mathcal{H}}} \leq \frac{\gamma_n}{2^{m-1}}\big\}.$$

If the event $\mathcal{T}_j^A(\gamma_n)$ occurs, then it must occur for a function f_j belonging to some S_m, so that we have a function f_j such that $\|f_j\|_n/\|f_j\|_{\mathcal{H}} \leq \frac{\gamma_n}{2^{m-1}} =: r_m$, and

$$\begin{aligned}\big|\frac{1}{n}\sum_{i=1}^n w_i f_j(x_{ij})\big| &> 2C\,\|f_j\|_{\mathcal{H}}\big\{\gamma_n^2 + \gamma_n \frac{\|f_j\|_n}{\|f_j\|_{\mathcal{H}}}\big\} \\ &\geq 2C\,\|f_j\|_{\mathcal{H}}\big\{\gamma_n^2 + \frac{\gamma_n}{2^m}\big\} \\ &\geq C\,\|f_j\|_{\mathcal{H}}\big\{\gamma_n^2 + r_m\big\},\end{aligned}$$

which implies that $Z_n(w; r_m) \geq C\big\{\gamma_n^2 + r_m\big\}$. Consequently, by union bound and the tail bound (44), we have

$$\begin{aligned}\mathbb{P}[\mathcal{T}_j^A(\gamma_n)] &\leq c_1 \sum_{m=1}^\infty \exp\big\{-c_2 n\gamma_n^2\big(1 + (\gamma_n/r_m)^2\big)\big\} \\ &= c_1 \sum_{m=1}^\infty \exp\big\{-c_2 n\gamma_n^2\big(1 + 2^{2m-2}\big)\big\} \\ &\leq c_1' \exp(-c_2 n\gamma_n^2).\end{aligned}$$

Case B: In this case, we define the sets

$$S_m := \big\{2^{m-1}\gamma_n \leq \frac{\|f_j\|_n}{\|f_j\|_{\mathcal{H}}} \leq 2^m\gamma_n\big\} \qquad \text{for } m = 1, 2, \ldots, M,$$

where $M = 2\log_2(1/\gamma_n)$ so that $2^M\gamma_n \geq 1$. By the same argument, we then have

$$\begin{aligned}\mathbb{P}[\mathcal{T}_j^B(\gamma_n)] &\leq M\ c_1 \exp(-c_2 n\gamma_n^2) \\ &\leq c_1 \exp\big(-c_2 n\gamma_n^2 + 2\log(1/\gamma_n)\big) \\ &\leq c_1' \exp(-c_2 n\gamma_n^2),\end{aligned}$$

by the condition $n\gamma_n^2 = \Omega(\log(1/\gamma_n))$.

B Proof of Lemma 2

Define the function

$$\widetilde{\mathcal{L}}(\Delta) := \frac{1}{2n}\sum_{i=1}^n \big(w_i - \Delta(x_i)\big)^2 + \lambda_n\|f^* + \Delta\|_{n,1} + \rho_n\|f^* + \Delta\|_{\mathcal{H},1}$$

and note that by definition of our M-estimator, the error function $\widehat{\Delta} := \widehat{f} - f^*$ minimizes $\widetilde{\mathcal{L}}$. From the inequality $\widetilde{\mathcal{L}}(\widehat{\Delta}) \leq \widetilde{\mathcal{L}}(0)$, we obtain

$$\frac{1}{2}\|\widehat{\Delta}\|_n^2 \leq \frac{1}{n}\sum_{i=1}^n w_i\widehat{\Delta}(x_i) + \lambda_n\sum_{j=1}^d\big\{\|f_j^*\|_n - \|f_j^* + \widehat{\Delta}_j\|_n\big\} + \rho_n\sum_{j=1}^d\big\{\|f_j^*\|_{\mathcal{H}} - \|f_j^* + \widehat{\Delta}_j\|_{\mathcal{H}}\big\}.$$

Now for any $j \in S^c$, we have

$$\|f_j^*\|_n - \|f_j^* + \widehat{\Delta}_j\|_n \ = \ -\|\widehat{\Delta}_j\|_n, \quad \text{and} \quad \|f_j^*\|_{\mathcal{H}} - \|f_j^* + \widehat{\Delta}_j\|_{\mathcal{H}} \ = \ -\|\widehat{\Delta}_j\|_{\mathcal{H}}.$$

On the other hand, for any $j \in S$, the triangle inequality yields

$$\|f_j^*\|_n - \|f_j^* + \widehat{\Delta}_j\|_n \leq \|\widehat{\Delta}_j\|_n,$$

with a similar inequality for the terms involving $\|\cdot\|_{\mathcal{H}}$. Since $\frac{1}{2}\|\widehat{\Delta}\|_n^2 \geq 0$, we conclude that

$$0 \leq \frac{1}{n}\sum_{i=1}^n w_i\widehat{\Delta}(x_i) + \lambda_n\big\{\|\widehat{\Delta}_S\|_{n,1} - \|\widehat{\Delta}_{S^c}\|_{n,1}\big\} + \rho_n\big\{\|\widehat{\Delta}_S\|_{\mathcal{H},1} - \|\widehat{\Delta}_{S^c}\|_{\mathcal{H},1}\big\}. \tag{47}$$

Recalling our conditioning on the event $\mathcal{T}(\gamma_n)$, we have the upper bound

$$\big|\frac{1}{n}\sum_{i=1}^n w_i\widehat{\Delta}(x_i)\big| \leq 2\,\kappa_3\ \big\{\gamma_n\|\widehat{\Delta}\|_{n,1} + \gamma_n^2\|\widehat{\Delta}\|_{\mathcal{H},1}\big\}.$$

Combining with the inequality (47) yields

$$\begin{aligned}0 &\leq 2\,\kappa_3\ \big\{\gamma_n\|\widehat{\Delta}\|_{n,1} + \gamma_n^2\|\widehat{\Delta}\|_{\mathcal{H},1}\big\} + \lambda_n\big\{\|\widehat{\Delta}_S\|_{n,1} - \|\widehat{\Delta}_{S^c}\|_{n,1}\big\} + \rho_n\big\{\|\widehat{\Delta}_S\|_{\mathcal{H},1} - \|\widehat{\Delta}_{S^c}\|_{\mathcal{H},1}\big\} \\ &\leq \frac{\lambda_n}{2}\|\widehat{\Delta}\|_{n,1} + \frac{\rho_n}{2}\|\widehat{\Delta}\|_{\mathcal{H},1} + \lambda_n\big\{\|\widehat{\Delta}_S\|_{n,1} - \|\widehat{\Delta}_{S^c}\|_{n,1}\big\} + \rho_n\big\{\|\widehat{\Delta}_S\|_{\mathcal{H},1} - \|\widehat{\Delta}_{S^c}\|_{\mathcal{H},1}\big\},\end{aligned}$$

where we have recalled our choices of (λ_n, ρ_n). Finally, re-arranging terms yields the claim (29).

C Proof of Lemma 3

Lemma 3 is a straightforward consequence of Lemma 7 from Appendix F. In particular, applying the latter lemma with $t = \gamma_n/2 \geq \epsilon_n$ yields

$$\|f_j\|_n \leq \|f_j\|_2 + \frac{\gamma_n}{2} \quad \text{for all } f_j \in \mathbb{B}_{\mathcal{H}}(1) \text{ and } \|f_j\|_2 \leq \gamma_n/2$$

with probability greater than $1 - c_1 \exp(-c_2 n\gamma_n^2)$. On the other hand, if $\|f_j\|_2 > \gamma_n/2$, then the sandwich relation (66) implies that $\|f_j\|_n \leq 2\|f_j\|_2$ with probability greater than $1 - c_1 \exp(-c_2 n\gamma_n^2)$. Defining the rescaled functions $g_j = 2f_j \in \mathbb{B}_{\mathcal{H}}(2)$, we have established that

$$\mathbb{P}[\mathcal{A}_j^c(\gamma_n)] \leq c_1 \exp(-c_2 n\gamma_n^2).$$

Recalling that $\mathcal{A}(\gamma_n) = \cap_{j=1}^d \mathcal{A}_j(\gamma_n)$, we can combine this upper bound with union bound, thereby obtaining

$$\mathbb{P}[\mathcal{A}^c(\gamma_n)] \leq d\, c_1 \exp(-c_2 n\gamma_n^2) \;\leq\; c_1 \exp(-c_2' n\gamma_n^2),$$

where we have used the fact that $\gamma_n = \Omega(\sqrt{\frac{\log d}{n}})$.

D Proof of Lemma 4

Define the alternative event

$$\mathcal{B}'(\delta_n) := \Big\{ \{\|h\|_n^2 \geq \delta_n^2/4 \quad \text{for all } h \in 2\mathcal{F} \text{ with } \|h\|_2 = \delta_n \Big\}.$$

We claim that it suffices to show that $\mathcal{B}'(\delta_n)$ holds with high probability. Indeed, given an arbitrary $g \in 2\mathcal{F} = \{f + f' \mid f, f' \in \mathcal{F}\}$ with $\|g\|_2 \geq \delta_n$, we can define $h = \frac{\delta_n}{\|g\|_2} g$. Since $g \in 2\mathcal{F}$ and $2\mathcal{F}$ is star-shaped, we have $h \in 2\mathcal{F}$, and also $\|h\|_2 = \delta_n$ by construction. Therefore, if $\mathcal{B}'(\delta_n)$ holds, we have $\|h\|_n^2 \geq \delta_n^2/4$, which implies that

$$\frac{\delta_n^2}{\|g\|_2^2} \|g\|_n^2 \;\geq\; \frac{\delta_n^2}{4},$$

or equivalently that $\|g\|_n^2 \geq \|g\|_2^2/4$, showing that $\mathcal{B}(\delta_n)$ holds.

Accordingly, the remainder of the proof is devoted to showing that $\mathcal{B}'(\delta_n)$ holds with high probability. For a truncation level $\tau > 0$ to be chosen, define the function

$$\phi_\tau(u) := \begin{cases} u^2 & \text{if } |u| \leq \tau \\ \tau^2 & \text{otherwise.} \end{cases} \tag{48}$$

By construction, ϕ_τ is continuous, Lipschitz with constant 2τ, and bounded by τ^2. Since $u^2 \geq \phi_\tau(u)$ for all $u \in \mathbb{R}$, we have

$$\frac{1}{n} \sum_{i=1}^n g^2(x_i) \geq \frac{1}{n} \sum_{i=1}^n \phi_\tau(g(x_i)). \tag{49}$$

The remainder of the proof consists of the following steps:

(1) First, we show that for all $g \in 2\mathcal{F}$ with $\|g\|_2 = \delta_n$, we have

$$\mathbb{E}[\phi_\tau(g(x))] \geq \frac{1}{2}\mathbb{E}[g^2(x)] \;=\; \frac{\delta_n^2}{2}. \tag{50}$$

(2) Next we prove that

$$\sup_{\substack{g \in \mathcal{F} \\ \|g\|_2 \leq \delta_n}} \Big|\frac{1}{n}\sum_{i=1}^n \phi_\tau(g(x_i)) - \mathbb{E}[\phi_\tau(g(x))]\Big| \leq \frac{\delta_n^2}{4} \tag{51}$$

with high probability.

Putting together the pieces, we conclude that for any $g \in \mathcal{F}$ with $\|g\|_2 = \delta_n$, we have

$$\frac{1}{n}\sum_{i=1}^n \phi_\tau(g(x_i)) \geq \frac{\delta_n^2}{2} - \frac{\delta_n^2}{4} \;=\; \frac{\delta_n^2}{4}$$

with high probability (to be specified later). Combined with the lower bound (49), this shows that event $\mathcal{B}'(\delta_n)$ holds with high probability, thereby completing the proof. It remains to establish the claims (50) and (51).

Establishing the lower bound (50): By the definition of ϕ_τ, we have

$$\begin{aligned}
\mathbb{E}\big[\phi_\tau(g(x))\big] &\geq \mathbb{E}\big[g^2(x)\,\mathbb{I}[|g(x)| < \tau]\big] \\
&= \mathbb{E}[g^2(x)] - \mathbb{E}\big[g^2(x)\,\mathbb{I}[|g(x)| \geq \tau]\big] \\
&= \delta_n^2 - \mathbb{E}\big[g^2(x)\,\mathbb{I}[|g(x)| \geq \tau]\big].
\end{aligned}$$

Consequently, it suffices to show that, with appropriate choice of the truncation level, we have $\mathbb{E}\big[g^2(x)\,\mathbb{I}[|g(x)| \geq \tau]\big] \leq \delta_n^2/2$. By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\big(\mathbb{E}\big[g^2(x)\,\mathbb{I}[|g(x)| \geq \tau]\big]\big)^2 &\leq \mathbb{E}[g^4(x)]\;\mathbb{E}\big[\mathbb{I}^2[|g(x)| \geq \tau]\big] \\
&= \mathbb{E}[g^4(x)]\;\mathbb{P}[|g(x)|^2 \geq \tau^2] \\
&\leq \mathbb{E}[g^4(x)]\;\frac{\delta_n^2}{\tau^2},
\end{aligned} \tag{52}$$

where the final step uses Markov's inequality, and the fact $\mathbb{E}[g^2(x)] = \delta_n^2$. It remains to bound the fourth moment. Any $g \in 2\mathcal{F}$ can be written as a sum $g = \sum_{j \in U} g_j$ of univariate functions over a subset U of cardinality at most $2s$, so that

$$\mathbb{E}[g^4(x)] = \mathbb{E}\big[\big(\sum_{j \in U} g_j(x_j)\big)^4\big] \;=\; \sum_{j \in U}\mathbb{E}[g_j^4(x_j)] + \binom{4}{2}\sum_{j \in U}\sum_{k \in U \setminus \{j\}}\mathbb{E}[g_j^2(x_j)]\;\mathbb{E}[g_k^2(x_k)],$$

where we have used the binomial expansion, the independence of x_j from co-ordinate to co-ordinate, and the fact that $\mathbb{E}[g_j(x_j)] = 0$. Re-arranging the second sum yields

$$\begin{aligned}\mathbb{E}[g^4(x)] &= \sum_{j \in U} \mathbb{E}[g_j^4(x_j)] + \binom{4}{2} \sum_{j \in U} \mathbb{E}[g_j^2(x_j)] \sum_{k \in U \setminus \{j\}} \mathbb{E}[g_k^2(x_k)] \\ &\leq \sum_{j \in U} \mathbb{E}[g_j^4(x_j)] + \binom{4}{2} \Big\{ \sum_{j \in U} \mathbb{E}[g_j^2(x_j)] \Big\} \mathbb{E}[g^2(x)] \\ &= \sum_{j \in U} \mathbb{E}[g_j^4(x_j)] + \binom{4}{2} \delta_n^4.\end{aligned}$$

For a function $g = \sum_{j \in U} g_j \in 2\mathcal{F}$, each univariate function satisfies $\|g_j\|_\infty \leq 2$, so that we have $\mathbb{E}[g_j^4(x_j)] \leq 4\mathbb{E}[g_j^2(x_j)]$, and hence $\sum_{j \in U} \mathbb{E}[g_j^4(x_j)] \leq 4 \sum_{j \in U} \mathbb{E}[g_j^2(x_j)] = 4\delta_n^2$. Overall, we have shown that

$$\mathbb{E}[g^4(x)] \leq 4\delta_n^2 + 6\delta_n^4 \leq 10\delta_n^2.$$

Substituting back into the inequality (52), we find that

$$\mathbb{E}\big[g^2(x)\, \mathbb{I}[|g(x)| \geq \tau]\big] \leq \sqrt{10\delta_n^2} \sqrt{\frac{\delta_n^2}{\tau^2}} \leq \frac{\sqrt{10}\, \delta_n^2}{\tau},$$

so that setting $\tau = 2\sqrt{10}$ is sufficient to prove the claim (50).

Establishing the bound (51): For a fixed subset U, recall the definition (4) of the function class $\mathcal{H}(U)$. Note that the function class $2\mathcal{F}$ can be written as $\bigcup_{|U|=2s} \mathcal{H}(U)$, where the co-ordinate functions g_j satisfy the bound $\|g_j\|_\infty \leq 2$. Accordingly, we define the random variable

$$Z_n(U) := \sup_{\substack{g \in \mathcal{H}(U) \\ \|g\|_2 \leq \delta_n}} \Big| \frac{1}{n} \sum_{i=1}^n \phi_\tau(g(x_i)) - \mathbb{E}[\phi_\tau(g(x))] \Big|, \tag{53}$$

and claim that it suffices to show that

$$\mathbb{P}\big[Z_n(U) \geq \frac{1}{16}(\delta_n^2 + t\,\delta_n + t^2)\big] \leq c_1 \exp(-c_2 n t^2) \qquad \text{for all } t > 0. \tag{54}$$

Indeed, assuming that this bound holds, then by applying the union bound over all $\binom{d}{2s}$ subsets of cardinality at most, we have

$$\mathbb{P}\Big[\sup_{\substack{g \in 2\mathcal{F} \\ \|g\|_2 \leq \delta_n}} \Big| \frac{1}{n} \sum_{i=1}^n \phi_\tau(g(x_i)) - \mathbb{E}[\phi_\tau(g(x))] \Big| \geq \frac{1}{16}(\delta_n^2 + t\,\delta_n + t^2)\Big] \leq c_1 \exp\Big\{ - c_2 n t^2 + \log \binom{d}{2s} \Big\}.$$

Setting $t = \delta_n$ and noting that our choice of δ_n ensures that $\frac{c_2}{2} n \delta_n^2 \geq \log \binom{d}{2s}$ yields the claim (51).

Accordingly, we now prove the bound (54). The functions $\phi_\tau(g(x))$ are uniformly bounded by τ^2. Moreover, since $\phi_\tau(u) = \min\{u^2, \tau^2\}$, we have

$$\mathbb{E}[\phi_\tau^2(g(x))] \leq \tau^2\, \mathbb{E}[\phi_\tau(g(x))] \leq \tau^2 \mathbb{E}[g^2(x)] \leq \tau^2 \delta_n^2,$$

where the final inequality uses the fact that $\mathbb{E}[g^2(x)] \leq \delta_n^2$. Consequently, we have shown that $\operatorname{var}(\phi_\tau(g(x))) \leq \mathbb{E}[\phi_\tau^2(g(x))] \leq \tau^2\delta_n^2$. We now apply Corollary 7.9 in Ledoux [19] with $\epsilon = 1$, $r = c_2 n t^2$ and $\sigma^2 = n\tau^2\delta_n^2$ to conclude that

$$\mathbb{P}\big[Z_n(U) \geq 2\,\mathbb{E}[Z_n(U)] + \frac{1}{16}t\delta_n + \frac{1}{16}t^2)\big] \leq c_1 \exp(-c_2 n t^2) \tag{55}$$

for some universal constants (c_1, c_2). (In this step, we can choose c_2 small enough so as to obtain the constants $1/16$.)

Based on the bound (55), our remaining task is to show that $\mathbb{E}[Z_n(U)] \leq \frac{1}{32}\delta_n^2$. By a standard symmetrization argument, we have

$$\mathbb{E}[Z_n(U)] \leq 2\mathbb{E}_{x,\sigma}\big[\sup_{\substack{g\in\mathcal{H}(U)\\ \|g\|_2\leq\delta_n}} \big|\frac{1}{n}\sum_{i=1}^n \sigma_i\phi_\tau(g(x_i))\big|\big],$$

where $\{\sigma_i\}_{i=1}^n$ is an i.i.d. sequence of Rademacher variables. Since the function ϕ_τ is Lipschitz with constant 2τ, the Ledoux-Talagrand contraction inequality (p. 112, [20]) implies that

$$\mathbb{E}[Z_n(U)] \leq 4\tau\,\mathbb{E}_{x,\sigma}\big[\sup_{\substack{g\in\mathcal{H}(U)\\ \|g\|_2\leq\delta_n}} \big|\frac{1}{n}\sum_{i=1}^n \sigma_i g(x_i)\big|\big]$$

Note that $\mathcal{H}(U)$ is subset of an RKHS with norm $\|g\|^2_{\mathcal{H}(U)} = \sum_{j\in U}\|g_j\|^2_{\mathcal{H}}$. Since $\|g_j\|_{\mathcal{H}} \leq 2$ for each g_j, we have $\|g\|_{\mathcal{H}(U)} \leq 4\sqrt{s}$ for all $g \in \mathcal{H}(U)$. Consequently, we have

$$\mathbb{E}[Z_n(U)] \leq 4\tau\;\mathbb{E}_{x,\sigma}\big[\sup_{\substack{\|g\|_{\mathcal{H}(U)}\leq 4\sqrt{s}\\ \|g\|_2\leq\delta_n}} \big|\frac{1}{n}\sum_{i=1}^n \sigma_i g(x_i)\big|\big].$$

Now defining the rescaled functions $h = g/\sqrt{s}$, we have

$$\begin{aligned}\mathbb{E}[Z_n(U)] &\leq 4\tau\;\sqrt{s}\;\mathbb{E}_{x,\sigma}\big[\frac{1}{n}\sup_{\substack{\|h\|_{\mathcal{H}(U)}\leq 4\\ \|h\|_2\leq\frac{\delta_n}{\sqrt{s}}}} \frac{1}{n}\sum_{i=1}^n \sigma_i h(x_i)\big] \\ &\leq \underbrace{32\tau\;\sqrt{s}\;\frac{1}{\sqrt{n}}\big[\sum_{\ell=1}^\infty \min\big\{\frac{\delta_n^2}{s}, \alpha_\ell\big\}\big]^{1/2}}_{T_n(\delta_n)}\end{aligned}$$

where $\alpha_1 \geq \alpha_2 \geq \cdots$ are the eigenvalues of the kernel associated with the Hilbert space $\mathcal{H}(U)$. This last inequality makes uses of standard upper bounds on kernel Rademacher complexities (e.g., see Mendelson [24]).

Now since $\mathcal{H}(U)$ is a sum of at most $2s$ copies of the same univariate Hilbert space $\mathcal{H}$, the eigenvalues $\{\alpha_\ell\}_{\ell=1}^\infty$ correspond to at most $2s$ copies of the eigenvalues $\{\mu_\ell\}_{\ell=1}^\infty$ of $\mathcal{H}$. Consequently, by factoring out these $2s$ terms, we obtain

$$T_n(\delta_n) \leq 64\tau\;s\;\frac{1}{\sqrt{n}}\big[\sum_{\ell=1}^\infty \min\{\frac{\delta_n^2}{s}, \mu_\ell\}\big]^{1/2}.$$

Now, as long as $\delta_n^2/s \geq \nu_n^2$, where ν_n is the critical rate (12) for the univariate kernel, we are guaranteed that

$$\frac{1}{\sqrt{n}}\big[\sum_{\ell=1}^{\infty}\min\{\frac{\delta_n^2}{s},\mu_\ell\}\big]^{1/2} \leq \frac{1}{\kappa_0}\,\frac{\delta_n^2}{s},$$

and hence $\mathbb{E}[Z_n(U)] \leq \frac{64\,\tau\,s}{\kappa_0}\,\frac{\delta_n^2}{s}$. Choosing $\tau = 2\sqrt{10}$ and $\kappa_0 = 64^2\sqrt{10}$, we conclude that $\mathbb{E}[Z_n(U)] \leq \frac{\delta_n^2}{32}$, as required.

E Proof of Lemma 5

Proof of part (a): Let $N = M(\frac{\delta}{\sqrt{s}}; \mathbb{B}_{\mathcal{H}}(1), \|\cdot\|_2) - 1$, and define $\mathcal{I} = \{0, 1, \ldots, N\}$. Using $\|u\|_0 = \sum_{j=1}^d \mathbb{I}[u_j \neq 0]$ to denote the number of non-zero components in a vector, consider the set

$$\mathfrak{S} := \big\{u \in \mathcal{I}^d \;\mid\; \|u\|_0 = s\big\}. \tag{56}$$

Note that this set has cardinality $|\mathfrak{S}| = \binom{d}{s}N^s$, since any element is defined by first choosing s co-ordinates are non-zero, and then for each co-ordinate, choosing non-zero entry from a total of N possible symbols.

For each $j = 1, \ldots, d$, let $\{0, f_j^1, f_j^2, \ldots, f_j^N\}$ be a $\delta/\sqrt{s}$-packing of $\mathbb{B}_{\mathcal{H}}(1)$. Based on these packings of the univariate function classes, we can use $\mathfrak{S}$ to index a collection of functions contained inside $\mathcal{F}$. In particular, any $u \in \mathfrak{S}$ uniquely defines a function $g^u = \sum_{j=1}^d g_j^{u_j} \in \mathcal{F}$, with elements

$$g_j^{u_j} = \begin{cases} f_j^{u_j} & \text{if } u_j \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{57}$$

Since $\|u\|_0 = s$, we are guaranteed that at most s co-ordinates of g are non-zero, so that $g \in \mathcal{F}$.

Now consider two functions g^u and h^v contained within the class $\{g^u, u \in \mathfrak{S}\}$. By definition, we have

$$\|g^u - h^v\|_2^2 = \sum_{j=1}^d \|f_j^{u_j} - f_j^{v_j}\|_2^2 \;\geq\; \frac{\delta^2}{s}\sum_{j=1}^d \mathbb{I}[u_j \neq v_j], \tag{58}$$

Consequently, it suffices to establish the existence of a "large" subset $\mathcal{A} \subset \mathfrak{S}$ such that the Hamming metric $\rho_H(u,v) := \sum_{j=1}^d \mathbb{I}[u_j \neq v_j]$ is at least $s/2$ for all pairs $u, v \in \mathcal{A}$, in which case we are guaranteed that $\|g - h\|_2^2 \geq \delta^2$. For any $u \in \mathfrak{S}$, we observe that

$$\Big|\big\{v \in \mathfrak{S} \;\mid\; \rho_H(u,v) \leq \frac{s}{2}\big\}\Big| \leq \binom{d}{\frac{s}{2}}\,(N+1)^{\frac{s}{2}}.$$

This bound follows because we simply need to choose a subset of size $s/2$ where u and v agree, and the remaining $s/2$ co-ordinates can be chosen arbitrarily in $(N+1)^{\frac{s}{2}}$ ways. For a given set $\mathcal{A}$, we write $\rho_H(u, \mathcal{A}) \leq \frac{s}{2}$ if there exists some $v \in \mathcal{A}$ such that $\rho_H(u,v) \leq \frac{s}{2}$. Using this notation, we have

$$\Big|\big\{u \in \mathfrak{S} \;\mid\; \rho_H(u,\mathcal{A}) \leq \frac{s}{2}\big\}\Big| \leq |\mathcal{A}|\,\binom{d}{\frac{s}{2}}\,(N+1)^{\frac{s}{2}} \overset{(a)}{<} |\mathfrak{S}|,$$

where inequality (a) follows as long as

$$|\mathcal{A}| \leq N^* \; := \frac{1}{2} \frac{\binom{d}{s}}{\binom{d}{\frac{s}{2}}} \; \frac{N^s}{(N+1)^{s/2}}.$$

Thus, as long as $|\mathcal{A}| \leq N^*$, there must exist some element $u \in \mathfrak{S}$ such that $\rho_H(u, \mathcal{A}) > \frac{s}{2}$, in which case we can form the augmented set $\mathcal{A} \cup \{u\}$. Iterating this procedure, we can form a set with N^* elements such that $\rho_H(u, v) \geq \frac{s}{2}$ for all $u, v \in \mathcal{A}$.

Finally, we lower bound N^*. We have

$$\begin{aligned} N^* &\overset{(i)}{\geq} \frac{1}{2} \big(\frac{d-s}{s/2}\big)^{\frac{s}{2}} \; \frac{(N)^s}{(N+1)^{s/2}} \\ &= \frac{1}{2} \big(\frac{d-s}{s/2}\big)^{\frac{s}{2}} \; N^{s/2} \big(\frac{N}{N+1}\big)^{s/2} \\ &\geq \frac{1}{2} \big(\frac{d-s}{s/2}\big)^{\frac{s}{2}} \; N^{s/2}, \end{aligned}$$

where inequality (i) follows by elementary combinatorics (see Lemma 5 in the paper [27] for details). We conclude that for $s \leq d/4$, we have

$$\log N^* = \Omega\big(s \log \frac{d}{s} + s \log M(\frac{\delta}{\sqrt{s}}; \mathbb{B}_{\mathcal{H}}(1), \|\cdot\|_2)\big),$$

thereby completing the proof of Lemma 5(a).

Proof of part (b): In order to prove part (b), we instead let $N = M(\frac{1}{2}; \mathbb{B}_{\mathcal{H}}(1), \|\cdot\|_2) - 1$, and then follow the same steps. Since $\log N = \Omega(m)$, we have the modified lower bound

$$\log N^* = \Omega\big(s \log \frac{d}{s} + sm\big),$$

Moreover, instead of the lower bound (58), we have

$$\|g^u - h^v\|_2^2 = \sum_{j=1}^{d} \|f_j^{u_j} - f_j^{v_j}\|_2^2 \; \geq \; \frac{1}{4} \sum_{j=1}^{d} \mathbb{I}[u_j \neq v_j] \; \geq \; \frac{s}{8}, \tag{59}$$

using our previous result on the Hamming separation. Furthermore, since $\|f_j\|_2 \leq \|f_j\|_{\mathcal{H}}$ for any univariate function, we have the upper bound

$$\|g^u - h^v\|_2^2 = \sum_{j=1}^{d} \|f_j^{u_j} - f_j^{v_j}\|_2^2 \; \leq \; \sum_{j=1}^{d} \|f_j^{u_j} - f_j^{v_j}\|_{\mathcal{H}}^2.$$

By the definition (56) of $\mathfrak{S}$, at most $2s$ of the terms $f_j^{u_j} - f_j^{v_j}$ can be non-zero. Moreover, by construction we have $\|f_j^{u_j} - f_j^{v_j}\|_{\mathcal{H}} \leq 2$, and hence

$$\|g^u - h^v\|_2^2 \leq 8s.$$

Finally, by rescaling the functions by $\sqrt{8}\,\delta/\sqrt{s}$, we obtain a class of N^* rescaled functions $\{\widetilde{g}^u, u \in \mathcal{I}\}$ such that

$$\|\widetilde{g}^u - \widetilde{h}^v\|_2^2 \geq \delta^2, \quad \text{and} \quad \|\widetilde{g}^u - \widetilde{h}^v\|_2^2 \leq 64\delta^2,$$

as claimed.

F Results on kernel classes

In this appendix, we collect some basic results about reproducing kernel Hilbert spaces, useful in our analysis. Let $\mathcal{H}$ be an RKHS of functions $f : \mathcal{X} \to \mathbb{R}$. Let $\{\sigma_i\}_{i=1}^n$ be an i.i.d. sequence of Rademacher variables, and let $\{x_i\}_{i=1}^n$ be an i.i.d. sequence of variables from $\mathcal{X}$, drawn according to some distribution $\mathbb{Q}$. For each $t > 0$, we define the local Rademacher complexities

$$\widehat{\mathcal{Q}}_n(t) \; := \mathbb{E}_\sigma\big[\sup_{\substack{\|g\|_n \le t \\ \|g\|_{\mathcal{H}} \le 1}} \frac{1}{n}\sum_{i=1}^n \sigma_i g(x_i)\big], \quad \text{and} \quad \mathcal{Q}_n(t) \; := \mathbb{E}_{x,\sigma}\big[\sup_{\substack{\|g\|_2 \le t \\ \|g\|_{\mathcal{H}} \le 1}} \frac{1}{n}\sum_{i=1}^n \sigma_i g(x_i)\big] \tag{60}$$

By results due to Mendelson [24], there are universal constants $c_\ell \le c_u$ such that for all $t^2 \ge 1/n$, we have

$$\frac{c_\ell}{\sqrt{n}}\big[\sum_{j=1}^\infty \min\{t^2, \mu_j\}\big]^{1/2} \;\le\; \mathcal{Q}_n(t) \;\le\; \frac{c_u}{\sqrt{n}}\big[\sum_{j=1}^\infty \min\{t^2, \mu_j\}\big]^{1/2}. \tag{61}$$

Conditionally on $\{x_i\}_{i=1}^n$, the same bounds hold for $\widehat{\mathcal{Q}}_n(t)$ with the population eigenvalues $\{\mu_\ell\}_{\ell=1}^\infty$ replaced by the eigenvalues $\{\widehat{\mu}_\ell\}_{\ell=1}^n$ of the kernel matrix defined by the n samples. We let ϵ_n and $\widehat{\epsilon}_n$ denote (respectively) the smallest solutions (of size at least $1/\sqrt{n}$) to the inequalities

$$\mathcal{Q}_n(\epsilon_n) \le \frac{\epsilon_n^2}{256}, \quad \text{and} \quad \widehat{\mathcal{Q}}_n(\widehat{\epsilon}_n) \le 256\widehat{\epsilon}_n^2. \tag{62}$$

These two quantities correspond to the critical rates derived from the population and empirical eigenvalues respectively. (Our scaling by 256 is for later theoretical convenience.)

Our first result relates the critical rates based on the population and empirical eigenvalues. Recall that $\gamma_n := \kappa_1 \max\big\{\nu_n, \sqrt{\frac{\log d}{n}}\big\}$.

Lemma 6. *We have*

$$\mathbb{P}\big[\widehat{\epsilon}_n \le \gamma_n\big] \ge 1 - c_1 \exp(-c_2\gamma_n^2). \tag{63}$$

This result is exploited at the start of Appendix A. In particular, combined with union bound, it implies that the event $\mathcal{D}(\gamma_n)$ holds with high probability, as claimed.

Our second result provides uniform control on the difference between the empirical $\|\cdot\|_n$ and population $\|\cdot\|_2$ norms over $\mathcal{H}$. In particular, for a radius $t \ge \epsilon_n$, we define the event

$$\mathcal{E}(t) := \big\{\sup_{\substack{g \in \mathbb{B}_{\mathcal{H}}(1) \\ \|g\|_2 \le t}} \big|\|g\|_n - \|g\|_2\big| \ge \frac{t}{2}\big\}. \tag{64}$$

Lemma 7. *Suppose that $\|g\|_\infty \le 1$ for all $g \in \mathbb{B}_{\mathcal{H}}(1)$. Then there exists universal constants (c_1, c_2) such that for any $t \ge \epsilon_n$,*

$$\mathbb{P}\big[\mathcal{E}(t)\big] \le c_1 \exp(-c_2 n t^2). \tag{65}$$

Moreover, for any $t \ge \epsilon_n$, we have

$$\frac{\|g\|_2}{2} \;\le\; \|g\|_n \;\le\; \frac{3}{2}\|g\|_2 \qquad \text{for all } g \in \mathbb{B}_{\mathcal{H}}(1) \text{ with } \|g\|_2 \ge t \tag{66}$$

with probability at least $1 - c_1 \exp(-c_2 n t^2)$.

F.1 Proof of Lemma 7

Our proof is based on the random variable

$$Y_n(t) := \sup_{g \in \mathbb{B}_{\mathcal{H}}(1),\, \|g\|_2 \leq t} \big| \|g\|_n^2 - \|g\|_2^2 \big|,$$

If the event $\mathcal{E}(t)$ occurs, then there exists some $g \in \mathbb{B}_{\mathcal{H}}(1)$ such that $|\|g\|_n - \|g\|_2| \geq \frac{t}{2}$, whence

$$\big| \|g\|_n^2 - \|g\|_2^2 \big| = \big| \|g\|_n - \|g\|_2 \big| \, \big(\|g\|_n + \|g\|_2 \big) \; \geq \; \frac{t^2}{4}.$$

Therefore, it suffices to establish the upper bound

$$\mathbb{P}\big[Y_n(t) \geq \frac{t^2}{4} \big] \leq c_1 \exp(-c_2 n t^2).$$

We first bound deviations above the expectation using concentration theorems for empirical processes [19]. The supremum of the variances is upper bounded by

$$\gamma^2(t) := \sup_{g \in \mathbb{B}_{\mathcal{H}}(1),\, \|g\|_2 \leq t} \frac{1}{n} \sum_{i=1}^n \operatorname{var}(g^2(x_i) - \|g\|_2^2) \; = \; \sup_{g \in \mathbb{B}_{\mathcal{H}}(1),\, \|g\|_2 \leq t} \mathbb{E}\Big[\big(g^2(x) - \|g\|_2^2 \big)^2 \Big],$$

using the i.i.d. nature of the samples $\{x_i\}_{i=1}^n$. Moreover, since the functions are uniformly bounded by 1, we have

$$\gamma^2(t) \leq 32\, \mathbb{E}\Big[\big(g(x) + \|g\|_2 \big)^2 \Big] \; \leq 64 t^2, \tag{67}$$

where the final inequality uses the fact that $\mathbb{E}[g^2(x)] = \|g\|_2^2 \leq t^2$. Consequently, applying Corollary 7.9 in Ledoux [19] with $\epsilon = 1$, $r = nt^2$ and $\sigma^2 = 64t^2$, we conclude that there are universal constants such that

$$\mathbb{P}\big[Y_n(t) \geq 2\, \mathbb{E}[Y_n(t)] + \frac{t^2}{20} \big] \leq c_1 \exp(-c_2 n t^2). \tag{68}$$

We now upper bound the mean. By a standard symmetrization argument, we have

$$\mathbb{E}[Y_n(t)] \;\; \leq \;\; 2\, \mathbb{E}_{x,\sigma}\Big[\sup_{g \in \mathbb{B}_{\mathcal{H}}(1), \|g\|_2 \leq t} \big| \frac{1}{n} \sum_{i=1}^n \sigma_i \, g^2(X_i) \big| \Big],$$

where $\{\sigma_i\}_{i=1}^n$ are i.i.d. Rademacher variables. Since $\|g\|_\infty \leq 1$ for all $g \in \mathcal{F}$, we may may apply the Ledoux-Talagrand contraction theorem ([20], p. 112) to obtain that

$$\mathbb{E}[Y_n(t)] \leq 8\, \mathbb{E}_{x,\sigma}\Big[\sup_{g \in \mathbb{B}_{\mathcal{H}}(1), \|g\|_2 \leq t} \big| \frac{1}{n} \sum_{i=1}^n \sigma_i g(x_i) \big| \Big] \; = \; 8\, \mathcal{Q}_n(t).$$

But by our choice (62) of ϵ_n and since $t \geq \epsilon_n$, we have have $\mathcal{Q}_n(t) \leq \frac{t^2}{256}$. Combined with the bound (68), we conclude that

$$Y_n(t) \leq 8 \frac{t^2}{256} + \frac{t^2}{20} \; \leq \; \frac{t^2}{4}$$

with probability at least $1 - c_1 \exp(-c_2 n t^2)$, as claimed.

Finally, let us prove the sandwich relation (66). For any $g \in \mathbb{B}_{\mathcal{H}}(1)$ with $\|g\|_2 \geq t \geq \epsilon_n$, we can define the function $h := \frac{t}{\|g\|_2} g$. Note that $h \in \mathbb{B}_{\mathcal{H}}(1)$ and $\|h\|_2 = t$, so that when the bound (65) holds, we have $\|h\|_2 - \frac{t}{2} \leq \|h\|_n \leq \|h\|_2 + \frac{t}{2}$ or equivalently, that

$$\frac{t}{2} \leq \frac{t}{\|g\|_2} \|g\|_n \leq \frac{3t}{2},$$

with probability at least $1 - c_1 \exp(-c_2 n t^2)$, which establishes the claim (66).

F.2 Proof of Lemma 6

For any $t > 0$, define the two random variables

$$\widehat{Z}_n(t) := \sup_{\substack{\|g\|_n \leq t \\ \|g\|_{\mathcal{H}} \leq 1}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(x_i), \quad \text{and} \quad Z_n(t) := \sup_{\substack{\|g\|_2 \leq t \\ \|g\|_{\mathcal{H}} \leq 1}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(x_i),$$

and observe that $\mathbb{E}_\sigma[\widehat{Z}_n(t)] = \widehat{\mathcal{Q}}_n(t)$ and $\mathbb{E}_{x,\sigma}[Z_n(t)] = \mathcal{Q}_n(t)$.

For any function with $\|g\|_n \leq t$, we have $\sum_{i=1}^n \operatorname{var}_\sigma(\sigma_i g(x_i)) = n\|g\|_n^2 \leq nt^2$. Consequently, applying the lower bound in Corollary 7.9 of Ledoux [19] with $r = c_2 n t^2$ and $\epsilon = 1/2$, we obtain

$$\mathbb{P}\big[\widehat{Z}_n(t) \leq \frac{1}{2}\widehat{\mathcal{Q}}_n(t) - t^2\big] \leq c_1 \exp(-c_2 n t^2). \tag{69}$$

Similarly, for any function with $\|g\|_2 \leq t$, we have $\sum_{i=1}^n \operatorname{var}_{\sigma,x}(\sigma_i g(x_i)) = n\|g\|_2^2 \leq nt^2$. Consequently, applying the upper bound in Corollary 7.9 of Ledoux [19] with $r = c_2 n t^2$ and $\epsilon = 2$, we obtain

$$\mathbb{P}\big[Z_n(t) \geq 2\mathcal{Q}_n(t) + t^2\big] \leq c_1 \exp(-c_2 n t^2). \tag{70}$$

Now suppose that $\|g\|_2 > t \geq \epsilon_n$. Then conditioned on the sandwich relation (66), we are guaranteed that $\|g\|_n > \frac{t}{2}$. Taking the contrapositive, we conclude that $\|g\|_n \leq \frac{t}{2}$ implies $\|g\|_2 \leq t$, and hence that

$$\widehat{Z}_n(t/2) \leq Z_n(t) \qquad \text{for all } t \geq \epsilon_n, \tag{71}$$

under the stated conditioning.

For any $t \geq \epsilon_n$, the inequalities (69), (70) and (71) hold with probability at least $1 - c_1 \exp(-c_2 n t^2)$. Conditioning on these inequalities, we can set $t = \gamma_n > \epsilon_n$, and thereby obtain

$$\begin{aligned}
\widehat{\mathcal{Q}}_n(\gamma_n) &\overset{(a)}{\leq} 2\widehat{Z}_n(\gamma_n) + 2\gamma_n^2 \\
&\overset{(b)}{\leq} 2Z_n(2\gamma_n) + 2\gamma_n^2 \\
&\overset{(c)}{\leq} 4\mathcal{Q}_n(2\gamma_n) + 8\gamma_n^2 \\
&\overset{(d)}{\leq} 128\gamma_n^2,
\end{aligned}$$

where inequality (a) follows from the bound (69), inequality (b) follows from the bound (71), inequality (c) follows from the bound (70), and inequality (d) follows since $2\gamma_n > \epsilon_n$ and the definition of ϵ_n. By the definition of $\widehat{\epsilon}_n$ as the minimal t such that $\widehat{\mathcal{Q}}_n(t) \leq 256 t^2$, we conclude that $\widehat{\epsilon}_n \leq \gamma_n$, as claimed.

References

[1] K. S. Alexander. Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probability Theory and Related Fields*, 75:379–423, 1987.

[2] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[3] F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[4] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. Submitted to Annals of Statistics, 2008.

[5] M. S. Birman and M. Z. Solomjak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Math. USSR-Sbornik*, 2(3):295–317, 1967.

[6] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.

[7] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37:373–384, 1995.

[8] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, pages 169–194, 2007.

[9] B. Carl and I. Stephani. *Entropy, compactness and the approximation of operators*. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, UK, 1990.

[10] B. Carl and H. Triebel. Inequalities between eigenvalues, entropy numbers and related quantities of compact operators in banach spaces. *Annals of Mathematics*, 251:129–133, 1980.

[11] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[12] Howard L. Weinert (ed.), editor. *Reproducing Kernel Hilbert Spaces : Applications in Statistical Signal Processing*. Hutchinson Ross Publishing Co., Stroudsburg, PA, 1982.

[13] C. Gu. *Smoothing spline ANOVA models*. Springer Series in Statistics. Springer, New York, NY, 2002.

[14] R. Z. Has'minskii. A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory Prob. Appl.*, 23:794–798, 1978.

[15] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986.

[16] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Jour. Math. Anal. Appl.*, 33:82–95, 1971.

[17] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *Proceedings of COLT*, 2008.

[18] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. Technical report, Georgia Tech., April 2010.

[19] M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001.

[20] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

[21] Y. Lin and H. H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, 34:2272–2297, 2006.

[22] P. Massart. About the constants in talagrand's concentration inequalities for empirical processes. *Annals of Probability*, 28(2):863–884, 2000.

[23] L. Meier, S. van de Geer, and P. Buhlmann. High-dimensional additive modeling. *Annals of Statistics*, 37:3779–3821, 2009.

[24] S. Mendelson. Geometric parameters of kernel machines. In *Proceedings of COLT*, pages 29–43, 2002.

[25] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209:415–446, 1909.

[26] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *NIPS Conference*, 2009.

[27] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. Technical Report arXiv:0910.2042, UC Berkeley, Department of Statistics, 2009.

[28] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. SpAM: sparse additive models. *Journal of the Royal Statistical Society, Series B*, 2010. To appear.

[29] S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Longman Scientific & Technical, Harlow, UK, 1988.

[30] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[31] C. J. Stone. Additive regression and other nonparametric models. *Annals of Statistics*, 13(2):689–705, 1985.

[32] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[33] A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY, 1996.

[34] G. Wahba. *Spline models for observational data*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, PN, 1990.

[35] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[36] B. Yu. Assouad, Fano and Le Cam. *Research Papers in Probability and Statistics: Festschrift in Honor of Lucien Le Cam*, pages 423–435, 1996.

[37] M. Yuan. Nonnegative garrote component selection in functional anova models. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pages 660–666, 2007.